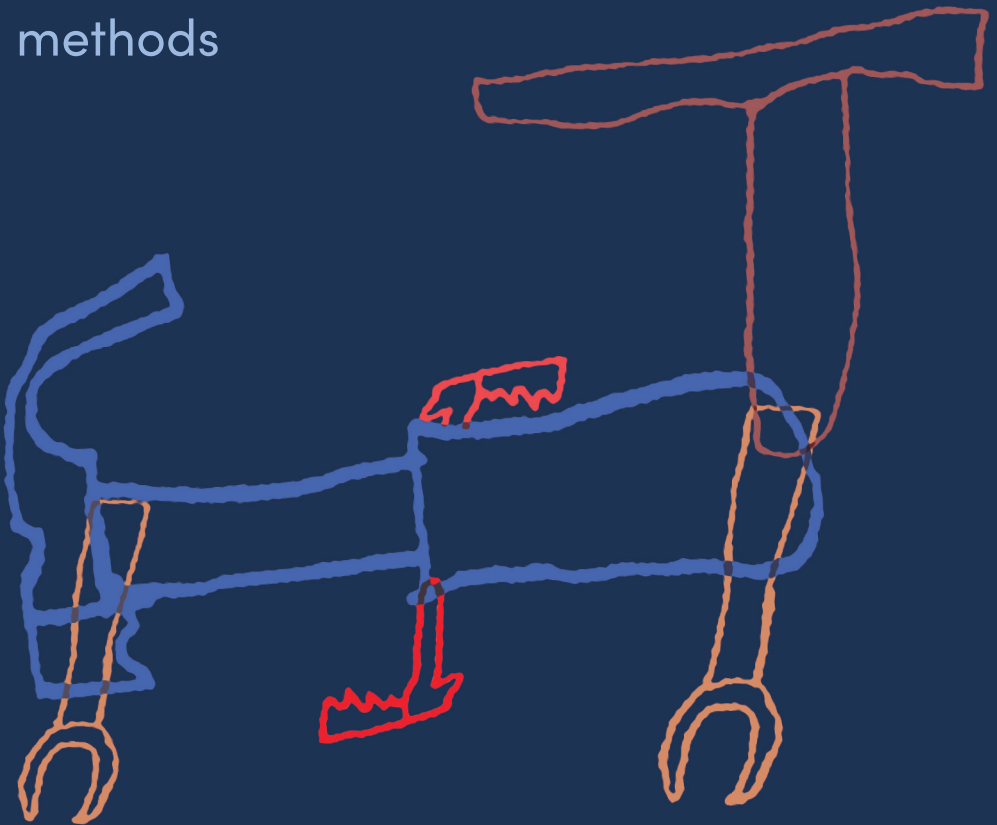


# Common Research Methods in Uncommon Applications

Replication research  
and qualitative  
methods



Hidde J. Leplaa



**Common Research Methods in Uncommon Applications:  
Replication research and qualitative methods**

Gebruikelijke Onderzoeksmethoden in Ongebruikelijke Toepassingen:  
Replicatiestudies en kwalitatieve methoden  
(met een samenvatting in het Nederlands)

**Proefschrift**

ter verkrijging van de graad van doctor aan de  
Universiteit Utrecht  
op gezag van de  
rector magnificus, prof. dr. ir. W. Hazeleger,  
ingevolge het besluit van het College voor Promoties  
in het openbaar te verdedigen op  
  
vrijdag 13 november 2026 des middags te 12.15 uur

door

**Hidde Jelmer Leplaa**

geboren op 27 april 1991  
te Lelystad

**Promotoren:**

Prof. dr. I.G. Klugkist

Prof. dr. H.J.A. Hoijsink

**Copromotor:**

Dr. C. Rietbergen

**Beoordelingscommissie:**

Prof. dr. R.H.H. Groenwold

Prof. dr. P.J. Lugtig

Dr. ir. M. Moerbeek

Prof. dr. A.G.J. van de Schoot

Prof. dr. M.A. Yerkes

Common Research Methods in Uncommon Applications: Replication research  
and qualitative methods

Proefschrift Universiteit Utrecht, Utrecht

Met samenvatting in het Nederlands.

Cover by: Erwin Timmerman

Printed by: ProefschriftMaken

ISBN: 978-90-393-8112-0

DOI: <https://doi.org/10.33540/3779>



©2026, Hidde Jelmer Leplaa

This work is licensed under a Creative Commons Attribution (CC BY) license.



*‘If all you have is a hammer,  
everything looks like a nail’*

- Maslow’s hammer

Maslow, in *The Psychology of Science* (1966)



## Contents

|          |                                                                                                     |           |
|----------|-----------------------------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                                                                                 | <b>9</b>  |
|          | Bibliography . . . . .                                                                              | 13        |
| <b>2</b> | <b>Replication according to psychology researchers</b>                                              | <b>17</b> |
| 2.1      | Introduction . . . . .                                                                              | 18        |
| 2.2      | Methods . . . . .                                                                                   | 20        |
| 2.3      | Results . . . . .                                                                                   | 23        |
| 2.4      | Discussion . . . . .                                                                                | 31        |
|          | Bibliography . . . . .                                                                              | 34        |
|          | Appendix 2.A: Topic list Replication in Psychology . . . . .                                        | 38        |
| <b>3</b> | <b>Bayesian evaluation of replication studies</b>                                                   | <b>45</b> |
| 3.1      | Introduction . . . . .                                                                              | 46        |
| 3.2      | The running example . . . . .                                                                       | 50        |
| 3.3      | What should be replicated . . . . .                                                                 | 51        |
| 3.4      | Conducting the replication study . . . . .                                                          | 62        |
| 3.5      | Example 2 - Janssen, Schirm, Mahon, and Caramazza . . . . .                                         | 69        |
| 3.6      | Discussion . . . . .                                                                                | 75        |
|          | Bibliography . . . . .                                                                              | 78        |
|          | Appendix 3.A: The Approximate Adjusted Fractional Bayes Factor . . . . .                            | 83        |
| <b>4</b> | <b>Applying qualitative methods to experimental designs: A tutorial for the behavioral sciences</b> | <b>89</b> |
| 4.1      | Context of the worked example . . . . .                                                             | 93        |
| 4.2      | Methodological approach . . . . .                                                                   | 94        |
| 4.3      | Discussion . . . . .                                                                                | 116       |
|          | Bibliography . . . . .                                                                              | 121       |

|                                                                                     |            |
|-------------------------------------------------------------------------------------|------------|
| <b>5 Prompting in practice: A qualitative field experiment on prompting empathy</b> | <b>127</b> |
| 5.1 Introduction . . . . .                                                          | 128        |
| 5.2 Method . . . . .                                                                | 130        |
| 5.3 Results . . . . .                                                               | 137        |
| 5.4 Discussion . . . . .                                                            | 145        |
| Bibliography . . . . .                                                              | 149        |
| Appendix 5.A: Observation list control week . . . . .                               | 153        |
| Appendix 5.B: Observation list intervention week . . . . .                          | 155        |
| Appendix 5.C: Topic list short interviews . . . . .                                 | 157        |
| Appendix 5.D: Topic list long interviews . . . . .                                  | 159        |
| <b>6 In conclusion</b>                                                              | <b>163</b> |
| Bibliography . . . . .                                                              | 165        |
| <b>Nederlandse samenvatting</b>                                                     | <b>169</b> |
| <b>About the author</b>                                                             | <b>173</b> |
| <b>Publications</b>                                                                 | <b>175</b> |
| <b>Dankwoord</b>                                                                    | <b>177</b> |

## CHAPTER 1

### Introduction

When still in high school, I saw science, scientists, and scientific research as the perfect situation. I imagined how smart people would come together to unite and solve some of the greatest questions and problems. At the start of my academic career, as an undergraduate student, my perspective on scientific research started to shift to a more realistic view.

The shift occurred when I started my bachelor's degree in psychology in 2010, with a specialization in social psychology. Throughout the program, we were regularly asked to formulate an interesting research question related to the course material. I consistently tried to choose questions that were not only relevant to the course material but also genuinely sparked my own curiosity. However, almost every first attempt was met with the same response from the instructors: "That is not a valid question for this course". It took a while before I grasped what was "wrong" with the research questions I formulated and considered most interesting. A lecturer explained: "This cannot be studied with the *quantitative methods* we use here. That is a *qualitative* research question". It was an epiphany for me. I realized that my natural instinct was to take a qualitative approach, whereas my bachelor's program mostly assumed a quantitative approach and design.

Qualitative research quickly became one of my personal interests. It generally uses a different philosophy to science than quantitative research. Typically, the quantitative approach sees the researcher as an outsider who tries to find an objective truth through observations, i.e. empirical data (Clark, 1998), without being part of the research situation. This can be labeled as a positivist philosophy (Michell, 2003), using the ontological view of realism (Pleasant, 2003). Most quantitative methods have in common that they value scientific approaches that are structured, in controlled situations, and can be repeated.

The qualitative approach sees the researcher as the essential tool to learn about and from the target population. The goal of a qualitative researcher is to put themselves in the shoes of their participants. This follows another philosophical approach to science which is frequently labeled as interpretative (Pervin & Mokhtar, 2022). Qualitative approaches often use a constructivist (Magoon, 1977) view on conducting science, using the ontological view of critical realism (Fleetwood et al., 2002). What most qualitative approaches have in common is that they believe that the social world is not only the product of objectively measurable phenomena, but that the social world is (also) the product of the minds of people. Zooming in on small facets of these phenomena, as is done with reductionism in many quantitative methods, is less valuable according to qualitative philosophies. Through directly learning from the participants, researchers would get a holistic understanding of the situation or topic they are interested in. It is exactly this quality of qualitative research methods that I investigated in the second part of this dissertation.

The first part of the dissertation was however inspired by another moment in my studies that also reshaped my view on science. This second moment occurred because I was among the minority of students with a strong affinity for research methods. Where others feared the statistics courses or were intimidated by SPSS and other analytical tools, I actually enjoyed them. I appreciated how it stimulated critical thinking about what empirical data can and cannot teach us about the world around us.

In one of the courses I took, we were tasked with analyzing a dataset. Through clear instructions and support from the supervisor I smoothly went through the assignment and had some time to spare. The instructor then invited me to do additional analyses, with questions like “Perhaps we could operationalize some variables differently?”, “What if we reduce the categories from five to three?” and “What if we select a different set of co-variates?”. I felt uncomfortable with these questions. The assignment started with a literature review, from which research questions were derived and operationalized into testable hypotheses. An analysis plan prescribed which analyses to perform and the results of these analyses were reported and interpreted. Doing more analyses would be highly exploratory and it was not clear to me what the purpose of them was within this assignment. The instructor motivated me to further explore the data in order to “get everything out of the data”, which felt unjust to me and as I later learned sounded very much like ‘fishing’ or *p*-hacking.

In this same period, the so-called replication crisis emerged, initiated by a publication by the Open Science Collaboration (OSC) that showed that many

psychological studies could not be replicated successfully. This means that when an original study is repeated in a highly similar way but with new data, the results and conclusions are not the same. This is not surprising occasionally but the OSC paper painted a disturbing picture. Their study concerned 100 original studies across different areas of psychology and concluded that only 35 led to similar results (Open Science Collaboration, 2012, 2015).

The publications of the OSC had its effect on the scientific community. It led to new procedures and guidelines, such as the increased use of preregistration (Lakens, 2024; Nosek et al., 2018) and registered reports (Lakens et al., 2024; Reich, 2021). Furthermore, the publications sparked other methodological research on the existence of the replication crisis. Researchers delved into the state of the behavioral sciences. For instance, multiple researchers studied the “researcher degrees of freedom”, i.e. the many decisions researchers have to make during a research process. Silberzahn et al. gave the same data set and research question to twenty-nine teams, resulting in twenty-nine different approaches and thus variance in the answers to the research question (Silberzahn et al., 2018). Another approach to this topic was taken by Steegen et al. who showed how performing all reasonable analyses instead of just a single one would identify the robustness of the results (Steegen et al., 2016). In addition, there have been “many labs”-projects in which the same study was conducted simultaneously across multiple sites and with multiple researchers (Forsell et al., 2019; Klein et al., 2014). These are some examples of recent studies contributing to a better understanding of the scientific world. Two studies in this dissertation were inspired by the replication crisis and the debates and papers that followed.

All in all, the result is a dissertation consisting of two parts. As said, the first two chapters were inspired by the replication crisis, which we studied in a novel way by applying Bayesian informative hypothesis testing to replication studies. In addition, we interviewed psychology researchers to add their points of view to those of methodologists. After these two studies were finished, COVID impacted both the world and this dissertation as it offered the opportunity to apply qualitative methods in a context where they are seldom used, that is to investigate experimental effects.

**Chapter 2** reports on a study from 2019. Publications by the OSC at that point had contributed to the emergence of what is widely referred to as the replication crisis. In the years after these publications, a substantial body of literature on replication research has been produced, primarily by methodologists and statisticians. However, comparatively little attention has been paid to the perspectives of researchers working within the field of psychology itself.

This chapter aims to address this gap by documenting the views of research in psychology on replication. Drawing on 11 semi-structured interviews with researchers from a psychology department in the Netherlands, we examined why replication studies are conducted and how they are experienced and practiced in everyday research contexts. Based on these data, we developed a typology consisting of four distinct perspectives on replication studies. These perspectives differ across ten characteristics, including the researcher's views on the goals of replication, methodological approaches, analytical strategies, and broader conceptions of science and its challenges. The resulting typology reflects key themes identified in the existing literature while also providing novel insights into how researchers in psychology interpret and engage with replication research. In doing so, this study has generated valuable insight into the perspectives of researchers within the disciplines of psychology.

In **Chapter 3** a method is proposed to determine whether the result from an original study is corroborated in a replication study. The paper is illustrated using two replication studies and the corresponding original studies from the reproducibility project psychology by the OSC (2012, 2015). The method we developed emphasizes the need to determine what one wants to replicate from the original paper. This can be done by translating the research hypotheses formulated in the introduction into informative hypotheses (Hoijsink, 2012), or, by translating the results into interval hypotheses (Lakens et al., 2013). It should offer an alternative to, among others, vote counting and null inferring (Anderson & Maxwell, 2016), to give a result that closely matches the question of the researchers. The Bayes factor will be used to determine whether the hypotheses resulting from the original study are corroborated by the replication study. Our method of assessing the successfulness of replication (1) allows users to define what aspects of the results matter most, (2) is straightforward to apply, and (3) focuses on the substantively meaningful outcomes of a study.

In **Chapter 4 and Chapter 5** we investigated the use of a qualitative approach in the context of a field experiment, to explore and demonstrate how alternative methods can provide valuable information that would not be obtained with the more standard approach. To be more specific, these chapters are about evaluating an experimental design with a qualitative approach. Studies with experimental designs are almost invariably evaluated with quantitative outcomes and methods, both in behavioral sciences and other disciplines. We argue that there can be added value of using qualitative methods for the evaluation of (behavioral) experiments. Incorporating qualitative data can enhance the ecological validity of a study, by acquiring a more holistic understanding of

the phenomenon of interest. There is, however, little methodological guidance on how to implement such an approach.

In **Chapter 4** we present methodological guidelines for the qualitative evaluation of results in experimental designs. Methodological guidelines are offered for each stage of a study, from formulation of the research goals, through data collection and data analysis, to the interpretation of a potential effect of the intervention. In addition, there is ample attention for ensuring the rigor of the research. The presented guidelines are developed and illustrated using an empirical example, in which a constructivist grounded theory approach was applied to evaluate the effect of empathy prompts on the motivation to adhere to COVID-19 regulations. The relationship between prompts and behavioral outcomes has been studied in clinical trials and quantitative research, but qualitative methods may offer deeper insight into underlying mechanisms.

**Chapter 5** presents this qualitatively evaluated experiment itself. We examined how empathy-inducing prompts influenced behavior, by collecting data during the COVID pandemic. Using an A-B design, the study alternated between control (A) and experimental (B) weeks, employing prompts to induce empathy during experimental weeks. Data came from predominantly white, highly educated students, collected through systematic and unsystematic observations complemented by short (Wave 1) and long (Wave 2) interviews. Findings revealed the complexity of student behavior, showing that while attitudes toward COVID-19 restrictions shaped intended behavior, these intentions did not consistently translate into actions. Contextual factors further influenced this relationship. Moreover, minimal differences between control and experimental groups suggested that the prompts had little measurable impact on behavioral outcomes.

In conclusion, the research presented in this thesis aims to offer a contribution to science, by investigating two themes: 1) replication research and how it is or could be used in psychology; 2) the added value of a qualitative approach, specifically in a context where it is rather uncommon (behavioral experiments). I hope that it triggers researchers to explore methodological pathways they did not consider before and that this provides additional value to (the rigor of) their research.

## Bibliography

- Anderson, S., & Maxwell, S. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21, 1–12. <https://doi.org/10.1037/met0000051>
- Clark, A. (1998). The qualitative-quantitative debate: Moving from positivism and confrontation to post-positivism and reconciliation. *Journal of Advanced Nursing*, 27, 1242–1249. <https://doi.org/10.1046/j.1365-2648.1998.00651.x>
- Fleetwood, S., Brown, A., & Roberts, J. (2002). The marriage of critical realism and marxism: Happy, unhappy or on the rocks? In *Critical realism and marxism*. Taylor; Francis. <https://doi.org/10.4135/9781483348858.n7>
- Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., & Dreber, A. (2019). Predicting replication outcomes in the many labs 2 study [Replications in Economic Psychology and Behavioral Economics]. *Journal of Economic Psychology*, 75, 102117. <https://doi.org/https://doi.org/10.1016/j.joep.2018.10.009>
- Hoijsink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. New York: Chapman; Hall/CRC.
- Klein, R., Ratliff, K., Vianello, M., R.B. Adams Jr., R., Bahnik, S., Bernstein, M., Brandt, M., Ijzerman, H., Bocian, K., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., ... Nosek, B. A. (2014). Data from investigating variation in replicability: A "many labs" replication project. *Journal of Open Psychology Data*, 2(1). <https://doi.org/10.5334/jopd.ad>
- Lakens, D., Scheel, A., & Isager, P. (2013). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1, 259–269. <https://doi.org/10.1177/2515245918770963>
- Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology*, 10(1), 117094. <https://doi.org/10.1525/collabra.117094>

- Lakens, D., Mesquida, C., Rasti, S., & Ditroilo, M. (2024). The benefits of preregistration and registered reports. *Evidence-Based Toxicology*, 2(1), 2376046. <https://doi.org/10.1080/2833373X.2024.2376046>
- Magoon, A. (1977). Constructivist approaches in educational research. *Review of Educational Research*, 47(4), 651–693. <https://doi.org/10.3102/00346543047004651>
- Michell, J. (2003). The quantitative imperative: Positivism, naïve realism and the place of qualitative methods in psychology. *Theory Psychology*, 13(1), 5–31. <https://doi.org/10.1177/0959354303013001758>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6), 657–660. <https://doi.org/10.1177/1745691612462588>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, 6251. <https://doi.org/10.1126/science.aac4716>
- Pervin, N., & Mokhtar, M. (2022). The interpretivist research paradigm: A subjective notion of a social context. *International Journal of Academic Research in Progressive Education and Development*, 11(2), 419–428. <https://doi.org/10.6007/IJARPED/v11-i2/12938>
- Pleasant, N. (2003). A philosophy for the social sciences: Realism, pragmatism, or neither? *Foundations of Science*, 8, 69–87. <https://doi.org/10.1023/A:1022441619760>
- Reich, J. (2021). Preregistration and registered reports. *Educational Psychologist*, 56(2), 101–109. <https://doi.org/10.1080/00461520.2021.1900851>
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Rosa, A. D., Dam, L., Evans, M. H., Cervantes, I. F., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3), 337–356. <https://doi.org/10.1177/2515245917747646>
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis [PMID: 27694465]. *Perspec-*

*tives on Psychological Science*, 11(5), 702–712. <https://doi.org/10.1177/1745691616658637>

## CHAPTER 2

### Replication according to psychology researchers

Leplaa, H.J., Rietbergen, C., & Klugkist, I. (2019). *Replication according to psychology researchers* [Unpublished manuscript]. Department of Methodology & Statistics, Utrecht University. Available since 2026 from: [https://osf.io/preprints/psyarxiv/ckxwr\\_v1](https://osf.io/preprints/psyarxiv/ckxwr_v1)

#### Abstract

Publications by the Open Science Collaboration have contributed to the emergence of what is now widely referred to as the replication crisis in psychology. In the years that followed, a substantial body of literature on replication research has been produced, primarily by methodologists and statisticians. However, comparatively little attention has been paid to the perspectives of researchers working within the field itself. This paper aims to address this gap by documenting their views on replication. Drawing on 11 semi-structured interviews with researchers from a psychology department in the Netherlands, we examined their views on why, how and when replications studies should be executed. Based on the data, we developed a typology consisting of four distinct perspectives on replication studies. These perspectives differ across ten characteristics, including views on the goals of replication, methodological approaches, analytical strategies, and broader conceptions of science and its challenges. The resulting typology reflects key themes identified in the existing literature while also providing novel insights into how researchers in psychology interpret and engage with replication research. In doing so, this study has generated insight into the perspectives of researchers within the disciplines of psychology.

*Keywords:* Replications, Social sciences, Replication Crisis

---

Author contribution: HL was involved in the initial research design, data collection, and wrote and revised the article. CR wrote and revised the article. IK was involved in data collection, and wrote and revised the article.

## 2.1 Introduction

The publication of the Open Science Collaboration (OSC) Reproducibility Project in Psychology, first announced (OSC, 2012) and later published (OSC, 2015), sent shock waves through the social sciences. Although it was not the first work to raise concerns about the replication crisis (Cohen, 1994, Tryon, 1992) or publishing practices in psychology (Rosenthal, 1979), the OSC project marked a turning point. Its unprecedented scale and public visibility brought reproducibility issues to the forefront of both scientific and public debate, reshaping discussions about research integrity, methodological rigor, and transparency in behavioral sciences (e.g., Ioannidis, 2018; Maxwell et al., 2015; Pashler & Wagenmakers, 2012; Stroebe et al., 2012; Szucs & Ioannidis, 2017). Against this backdrop, the present study conducted in 2019 aims to examine how researchers working within the field of psychology view the current state of replication within the social sciences.

In the years following the OSC project, a broad range of papers has been published examining the state of the social sciences, with a particular focus on replication. Replication refers to the repetition of a study with the aim of determining whether conducting it again leads to the same conclusions as the original work—that is, whether the findings can be obtained once more (Nosek & Errington, 2020). Within the diverse literature on replication, we identify three broad lines of discussion. First, various papers focus on the circumstances under which replication should be pursued. These contributions emphasize that replication, while undoubtedly important, should not be treated as an end in itself but rather as an element within a broader framework of scientific progress. For example, Finkel et al. (2017) argued that replicability constitutes only one aspect of good scientific practice. In their view, prioritizing replication studies can inadvertently divert attention from other essential aspects of research, such as discovery, internal validity, external validity, and consequentiality. In line with this reasoning, other authors have cautioned against performing replications indiscriminately, instead advocating replication efforts that are theoretically motivated and methodologically meaningful (Field et al., 2019).

Second, a line of discussion concentrates on methodological frameworks for conducting replication research, with particular emphasis on the typology of replication designs (e.g., Bettis et al., 2016; Brandt et al., 2015; Hüffmeier et al., 2016; Lynch et al., 2015). Replications are conventionally distinguished as *direct* or *conceptual* (Hüffmeier et al., 2016). Direct replications seek to reproduce the original investigation as faithfully as possible (Schmidt, 2009),

while conceptual replications intentionally modify aspects of the design while continuing to test the same underlying theoretical proposition (Schmidt, 2009; Stroebe & Strack, 2014). The principal contribution of direct replications lies in their ability to evaluate the robustness of empirical results, recognizing that even the most rigorous attempts will necessarily involve minor deviations in setting, sampling units, and materials or procedures (Cesario, 2014; Stroebe & Strack, 2014). Conceptual replications, on the contrary, serve to assess the generality of theoretical claims. As Crandall and Sherman argue, “if the idea replicates across operationalizations, it is substantially more likely to be correct than if it only replicates using the exact same operationalizations” (Crandall & Sherman, 2016). In demonstrating that theoretical constructs hold across varied operational and contextual conditions, conceptual replications strengthen the external validity and theoretical relevance of psychological phenomena (LeBel & Peters, 2011; Pashler & Harris, 2012). Collectively, these methodological approaches delineate a continuum that ranges from the assessment of empirical reliability to the evaluation of theoretical generalizability.

Building on these methodological distinctions, a third theme in the replication literature addresses the analysis of replication outcomes (e.g., Anderson & Maxwell, 2016; Etz & Vandekerckhove 2016). In OSC (2012, 2015), multiple analytical strategies were employed to evaluate corroboration, including vote counting, subjective judgment, and meta-analytic synthesis of effect sizes. Anderson and Maxwell (2016) provided an extensive reflection on the relative merits and limitations of these approaches, as well as on the criteria that should inform their selection. For instance, vote counting corresponds closely to the notion of reproducibility—the capacity to obtain consistent patterns of significant or non-significant results across independent studies. In contrast, subjective assessments may be more suitable for situations where duplication of analytic results is infeasible. In such contexts, subjective evaluation enables the focus on whether the replication yields comparable theoretical inferences rather than identical statistical results. This analytical perspective is often linked to the use of a conceptual design, illustrating how both the replication design and the inferential strategy jointly determine the interpretation of empirical corroboration.

The academic conversation on replication continues to evolve. This development is reflected in a growing body of manuscripts that address issues such as publishing challenges (Nosek et al., 2012), the need for transparency and openness (Nosek et al., 2015; Wicherts et al., 2016), questionable research practices (Shrout & Rodgers, 2018), and the role of replications in the social sci-

ences (Zwaan et al., 2018). Together, these contributions illustrate an ongoing debate about the value, objectives, and preferred forms of replication within contemporary science. Across this body of work, the discourse has been shaped predominantly by methodologists and statisticians whose insights have guided much of the conceptual and technical discussion. In contrast, the perspectives of substantive researchers—those conducting empirical work within any of the fields of psychology—have received less attention. The present study seeks to address this by exploring and documenting how researchers working in psychology experience and perceive replication practices within their own domain in this period of change and debate.

To do this, we conduct a qualitative study based on in-depth interviews with psychology researchers, to examine the three overarching themes identified in the literature: the motivations behind psychological researchers' engagement in replication studies, the methodological approaches they employ and the analytical strategies used to evaluate their findings. To enrich this evaluation, we explore the researcher's lived experiences in conducting replication studies. Insight into the experiences and ideas of researchers can elucidate why replications are (or are not) undertaken in particular ways and identify the motivations, barriers, and resources that shape these decisions. This inquiry could contribute to a more comprehensive understanding of the current state of replication in psychology. The remainder of this paper is organized as follows. In the methods section we outline the design of the study and interviews, the research participants and the planned analysis approach. Emerging themes from the data are presented in the results section, including descriptions of the different perspectives of the researchers. The paper is concluded with a discussion of the results.

## **2.2 Methods**

### **2.2.1 Data collection**

In this study a qualitative design was employed in which semi-structured interviews of approximately 45 minutes were conducted to collect the data between the summer of 2019 and the beginning of 2020. These interviews were conducted in meeting rooms or private offices in a university building. These are safe environments, isolated from other people, that contain a meeting room table and some chairs, as well as blank walls and a screen that was turned off during interviews.

A total of ten interviewers conducted the interviews, working in pairs. Within each pair, one interviewer served as the primary interviewer and led the conversation. The second interviewer was responsible for recording the interview, asking additional follow-up questions when necessary, and taking detailed notes. These notes captured both the content of the respondent's answers and reflective observations on the interview process. Furthermore, memos were taken throughout data collection to record moments of great impact. Some memos describe situations of possible steering by the interviewer, e.g. when their questions introduce a point of view which is adopted by the interviewee.

To ensure a similar procedure throughout all interviews, an instruction meeting was held to discuss how to approach the interviews. During the interviews, a topic list was used that included some starting questions and some optional follow-up questions. The topic list contained the following topics based on the three broad lines of discussion present throughout the replication literature: their work and specifically their experiences with replications, motivations to conduct replication studies, what they consider to be replications, how replications should be conducted and how the outcomes should be analyzed. Each topic was covered in the topic list with a starting question (e.g. '*When do you consider the outcomes of an original study to be replicated?*') and potential follow-up questions (e.g. '*What is the role of the authors of the original study?*', '*Should a replication study use the same measures as the original study?*'). In addition, three optional vignettes were added to the topic list to be used as eliciting material. These could be used as practical examples to further specify the views of the interviewee. In some cases, the interviewees gave enough examples themselves, and the vignettes were not introduced. The topic list with its vignettes is included in Appendix 2.A and can be found on OSF via <https://osf.io/s8hcy/>.

All interviews were transcribed verbatim, to remain close to the words of the participant (McMullin, 2023). As soon as the transcript was completed, it was sent to the participant to incorporate a member check into the study (Lincoln & Guba, 1985). They had two weeks to comment on the transcript if it seemed necessary. This resulted in small remarks and corrections, but no major revisions or retractions occurred.

### 2.2.2 Research participants

The participants in this study were all employed as researchers in the department of Psychology at Utrecht University. A sample matrix was used to ensure that

participants from different disciplines and with different levels of experience, i.e., in different career phases, were approached to participate in the research.

All participants were approached by mail to be part of a study with the work title “What is a replication study”. Reminders were sent after some time and some potential participants were asked in person if they were willing to participate. Participating would mean that they were willing to be interviewed for about an hour. In total, 22 psychology researchers refused to participate in the study or did not respond to the invitation or any of the reminders. The final sample consisted of six female and five male participants. Seven of them were Dutch natives. The distribution over disciplines and career phases is reported in Table 2.1.

**Table 2.1:** Number of participants per discipline and career phase.

|                                | Clinical<br>Psychology | Experimental<br>Psychology | Developmental<br>Psychology | Social and<br>Health Psychology |
|--------------------------------|------------------------|----------------------------|-----------------------------|---------------------------------|
| PhD-candidates<br>& Post-docs  | 1                      | 0                          | 1                           | 0                               |
| Assistant professors           | 2                      | 0                          | 2                           | 2                               |
| Associate &<br>full professors | 0                      | 1                          | 1                           | 1                               |

### 2.2.3 Ethical considerations

Ethical approval was acquired from the Ethics Committee of the Faculty of Social and Behavioral Sciences of Utrecht University in November 2018. The participants were aware of the research procedure and willingly and voluntarily participated in the study. Before conducting the interviews, an informed consent was signed.

### 2.2.4 Analysis

As transcripts were ready for analysis, they were uploaded to NVivo12 (2020). Initially, the data were intended to be analyzed following the procedures of the Grounded Theory approach (Glaser, 1978), using open, axial, and selective coding. Open coding was conducted first and was carried out independently by two researchers. Examples of open codes formulated during this phase include *uses p-value to determine replication*, *replications should have an important role in the field*, and *methods should be as similar as possible*. Theoretical saturation was reached at the level of open coding: after seven interviews had been coded, interviews eight and nine did not yield additional unique insights. Thus, the

last two interviews were not open coded, but they were used during the next stages of analysis.

Following open coding, one of the researchers initiated axial coding individually. During this phase, but before axial coding was completed, the idea of developing a typology emerged. Axial coding was continued with the aim of creating an initial code tree consisting of themes recurring across most of the data. This code tree subsequently served as the starting point for the development of the typology. During the development of the typology, the themes were re-examined and further elaborated into characteristics of researchers' views related to replication. Through constant comparison (Kolb, 2012), that is, the iterative process of moving back and forth between the data and the emerging typology, four types of researchers were identified with these expressions. This iterative process ultimately resulted in the typology presented in the results section. The data package is made available via <https://osf.io/s8hcy/>.

## 2.3 Results

In this section, we first report ten characteristics of researchers' perspectives on replication studies, which form the first column of the typology in Table 2.2. The characteristics are presented in *italics* in the following subsection. The second subsection describes the four types of researchers that we recognized in our sample and how their distinct perspectives on replication research can be described in terms of the ten characteristics.

### 2.3.1 Characteristics

The first characteristic is the researcher's *attitude toward replication*. This describes the value researchers give to replication studies and ranges from seeing them as essential at one extreme to seeing it as redundant at the other end. The attitude towards replication relates to what researchers believe should be the *primary goal of replication*. This characteristic entails what the interviewees mention as motivations to conduct replication studies, e.g. to test the reliability of findings or testing under which conditions a theory holds. This view in turn relates to a preference for *what to replicate*: only new insights, every study, or a preference to abandon replications altogether. Furthermore interviewees have different opinions on the *role of the original researcher* in a replication study. Some argued that the original researcher should not be involved in the replication study, while others advocated that they should be consulted in the process

or that they should collaborate with replicating researchers.

The fifth characteristic present throughout the data describes the *preferred methodology* for performing replications, varying from a preference for only performing conceptual replication to a preference for exclusively performing direct replications. The preference for a specific methodology is related to the *evaluation criteria* that researchers believe should be used. Although some researchers prefer the  $p$ -value, others formulate a broader definition of “predetermined measures”, or prefer a qualitative interpretation to determine if the results are corroborated. Some researchers even indicate that they would prefer to interpret and report the results as any normal study, without focusing on the goal of replication per se.

The first six characteristics presented flow logically from the topic list, which was designed based on topics prominent in the literature. Four more characteristics emerged organically during the interviews and relate to the view on science and the feasibility of replication research.

First, the interviewees mention different *views on the practice of science*. This characteristic describes the way researchers believe science works in practice, the overarching goals they pursue, and the ways in which these goals should be achieved. While some see science as a theory-building effort, either collaborative or not, others view science as more rule-governed and structured endeavor, or a process constrained by reality. The *views on the practice of science* are grounded in assumptions about science as an abstract phenomenon, as described in their *underlying beliefs about science*. Some explain science as a partly flawed phenomenon shaped by systematic barriers and limitations. Others emphasize the need for repeated systematic testing to achieve trust in theories through collective input. Some would argue for more creative ways to make progress rather than continuous repetition.

The beliefs about science link to *views on replication feasibility*, with a spectrum going from feasible and necessary to unnecessary and a waste of resources. A few researchers see limitations and barriers, which convinces some to move away from replication at a larger scale, yet others still view replication is desirable and should be aimed for. This characteristic links to some of the experienced *problems in science*. Some researchers express the opinion that science does not have any overarching problems. Other researchers do experience recurring problems in science, such as lack of funds and interest as some of the barriers to conduct replication, while others describe problems such as the file drawer phenomenon, questionable research practices, and a plethora of procedures and requirements.

### 2.3.2 Typology

Table 2.2 presents the typology that depicts four types that differ based on their attitude towards replication. The four types are described in terms of the ten characteristics introduced in the previous section.

#### Collectivist

The first type of researcher we identified conceptualizes science as a collective endeavor. From this perspective, the primary goal is to collaboratively develop robust theories that address fundamental questions in the field. The people that have this view barely experience competition within their work-field with coworkers. These researchers focus on the challenge of collaboratively unraveling the truth. The challenges they experience are thus directed at misinformation in general and more specifically to methodological shortcomings and erroneous conclusions, including those arising from issues such as the *file drawer problem*. This view is related to the view of replication as a crucial filtering mechanism to distinguish reliable findings from those resulting from Type I errors or misinterpretation.

*‘I think we would really benefit from collaboration. Working together to identify patterns across fields, both in the Netherlands and globally. By collaborating, we can actively work toward building broader theories that we can all move toward.’ (P[articipant]1)*<sup>1</sup>

This cooperative vision of science seems to extend to the way replication studies are valued. It aligns with the advocacy for close collaboration with original authors to ensure methodological rigor and interpretative clarity. From this point of view, any study is worth replicating, as each replication contributes to the cumulative body of knowledge. More than being valuable, replication is seen as essential to the scientific process. However, this idealism is likely to encounter frustration due to the lack of institutional support and funding for replication work.

---

<sup>1</sup>All reported quotes in the Results section are transcribed intelligent verbatim to improve the readability (McMullin, 2023). During analysis verbatim transcribed versions were used.

**Table 2.2: Typology of four types of researchers with a particular outlook on replication studies.**

| <div> <div>Type</div> <div>Characteristic</div> </div> | Collectivist                                                    | Precisionist                                         | Skeptic                                                      | Innovator                                                      |
|--------------------------------------------------------|-----------------------------------------------------------------|------------------------------------------------------|--------------------------------------------------------------|----------------------------------------------------------------|
| <b>Attitude toward replication</b>                     | Essential and valuable for all studies                          | Necessary for every study; should be routine         | Useful only in select, high-stakes cases                     | Low value; sees most replications as redundant or unoriginal   |
| <b>Primary goal of replication</b>                     | Strengthening theories through collective understanding         | Testing reliability and minimizing chance findings   | Verifying only critical findings under limited conditions    | Developing new insights from familiar ideas                    |
| <b>What to replicate</b>                               | Everything, necessary for good science                          | Everything, always                                   | New insights                                                 | Nothing, only applying theories in new contexts                |
| <b>Role original researcher</b>                        | In collaboration with the original researchers                  | Use original researcher for consult                  | Original researcher is not involved                          | Original researcher is not involved                            |
| <b>Preferred methodology</b>                           | Slight preference for conceptual                                | Preference for direct                                | Only direct                                                  | Only conceptual                                                |
| <b>Evaluation criteria</b>                             | A qualitative interpretation of statistical results             | Predetermined measures (e.g. range of effect size)   | “Objective” measures (e.g. <i>p</i> -values)                 | Interpret and report like any normal study                     |
| <b>View on the practice of science</b>                 | Collaborative, theory-building effort                           | Rule-governed, structured, precision-driven          | Resource-aware, constrained by systematic limitations        | Innovation-driven, open-ended, theory-expanding                |
| <b>Underlying belief about science</b>                 | Progress emerges through collective verification and refinement | Trust is earned through repeated, systematic testing | Remains unclear, they feel limited by “the system”           | Progress stems from creativity, not repetition                 |
| <b>View on replication feasibility</b>                 | Desirable despite systemic barriers                             | Feasible and necessary, with proper structure        | Impractical on a large scale; system not built to support it | Unnecessary; better to invest in novel, context-sensitive work |
| <b>Problems in science</b>                             | Lack of funds and interest                                      | File drawer                                          | QRP’s, lack of funding, overdoing bureaucracy                | No overarching problems                                        |

In terms of methodology, these cooperative ideas tend to coexist with a slight preference for conceptual replications. Although this perspective does not reject direct replications, it appears to be associated with the belief that introducing modifications can add interpretive depth and help test the generalizability of findings. The overarching goal is to assess whether the theoretical conclusions of the original study hold under varying conditions.

*‘What is more interesting is conceptual replication. To understand the parameters and conditions under which the effect is found or involved.’ (P2)*

When evaluating replication outcomes from this viewpoint, there is a tendency to favor qualitative interpretative assessments over purely statistical comparisons. Given the emphasis on conceptual replications, exact duplication is neither expected nor required and the key question is whether the replication study supports the broader theoretical claims of the original work. Each study, whether confirmatory or not, is regarded as a contribution to an evolving understanding of the phenomena under investigation. This orientation appears to be more common in our sample among early-career researchers.

### **Precisionist**

The second type of researcher, which we label *precisionists*, emphasizes the importance of structure, precision, and methodological rigor in scientific practice. Similarly to the *collectivist*, the *precisionist* has a positive attitude towards replication. They both view replication as essential for good science, though their underlying reasoning is different. While the *collectivist* prefers qualitative interpretation and flexible procedures, the *precisionist* tends to value precision and structure. The preference for structure could stem from the belief that scientific progress relies on systematic and clear-defined procedures. Minimizing ambiguity and avoiding interpretive flexibility are linked to this view, both within individual studies and across research programs.

*‘I think this is not really feasible, but in an ideal world, one person would design the study and preregister how it would be conducted, including the analysis or analytical approach, as far as possible. Then someone else, who does not know what the conclusions should be, would analyze the data and see what the results are.’ (P3)*

The desire for structure tends to be linked to a preference for direct replications, as they are best suited to answer the central question: ‘Was the original conclusion valid or merely the result of chance?’ The preference for direct replication could come from the desire to verify the findings in conditions as similar as possible to the original study. At the same time, these views are often paired with the experience of practical barriers to exact replication, particularly the lack of detailed reporting in original publications. For this reason, the participation of the original authors in the replication process is encouraged. Original authors should provide methodological feedback similar to commenting on a pre-registration, clarifying whether potential differences in study design are likely to affect outcomes.

In line with a preference for precision, the views of the *precisionist* are linked to a preference for clear and pre-specified criteria for evaluating replication success. For example, researchers could use effect size intervals derived from the original study to determine whether the replicated results are consistent. More complex designs, such as multi-factorial ANOVAs, may require evaluating multiple outcomes separately, some of which might be corroborated while others are not. The systematic criterion-based approach reflects the idea that interpretative clarity is the key to scientific reliability.

The viewpoint of the *precisionist*, similar to the ideas of the *collectivist*, is often accompanied with a clear opinion on the scope of replication: every study warrants replication. A single study pointing to a claim might be a good starting point, but without other studies corroborating the results the original study’s results are meaningless. In other words, no single result should be taken at face value. Robust science is built through the accumulation of consistent findings in multiple studies, which makes the *file drawer phenomenon* one of the bigger challenges in science. From this perspective, wide-scale replication is not just beneficial; it is essential. In our sample, the characteristics of a *precisionist* were mostly observed in the mid-career researchers.

‘*Science is about progress, not about success. It’s about finding out how the world works. Success is gaining a bit more knowledge about the world and uncovering some facts.* (P4)

## Skeptic

Where both the *collectivists* and *precisionists* clearly value replication research, a third group of researchers holds a skeptical view towards the practical relevance of replication studies. Although they will acknowledge the importance of verifying scientific claims, the *skeptic* view, as we labeled it, is linked to doubts about the utility of conducting replications. They mostly doubt the feasibility of direct replications given current constraints in funding, time, and institutional incentives.

The *skeptic* view includes the idea that meaningful progress is more likely to arise from conducting new studies in novel contexts or with different populations. Although some might categorize new studies in new contexts as conceptual replication, the skeptic view defines replication more narrow. In their view, the broad classification of conceptual replication is rejected. Their logic is that if every new study qualifies as conceptual replication, the term becomes virtually meaningless: almost every study adopts existing theories in new contexts. In the *skeptic's* terminology, replication studies refer exclusively to direct replications, defined narrowly and precisely.

*‘For me, it means a complete repetition of the same study but with different subjects; the latter is important. Beyond that, I would say that a replication should be an identical study— with the same paradigm, the same measures, the same statistical procedures, and the same methodology. A complete replication, for me, really means complete.’ (P5)*

Although within this view there is recognition for the theoretical value of replication, it is also seen as a luxury. In an ideal world with unlimited resources, replication would be a regular feature of scientific practice. However, under real-world constraints, the belief is that replications should be limited to specific, high-stakes scenarios, such as verifying novel claims before they inform policy or practice. This view is pragmatic: replication should serve a clear and consequential purpose, not be pursued for its own sake.

*‘I would be more convinced if there were studies using different outcome measures for depression and all of them found that the medication was effective. That would be more interesting to me than a single direct replication, and it sounds more convincing to me.’ (P6)*

Within the view of the *skeptic*, there seems to be a preference for objective quantitative criteria to evaluate replication outcomes, similar to the preferences of the *precisionist*. The *p*-value remains a preferred indicator, though there is room for awareness of its limitations and an acceptance of the use of alternative metrics, provided that they are clearly defined. The *skeptic's* view is characterized as a reserved attitude toward the relevance of (direct) replication efforts. *Skeptics* were found among the senior researchers in our sample.

### **Innovator**

The fourth and final type, which we refer to as the *innovator*, approaches the topic of replication with hesitation and disinterest. In our sample only one *innovator* was identified, this was a senior researcher. Similar to the *skeptic*, the *innovator* contrast the ideas from the *collectivist* and the *precisionist* who both value replication very highly. Both the *skeptic* and the *innovator* are more hesitant, though they see some uses for replication. Whereas the *skeptic* sees some use for direct replication, the *innovator* would prefer conceptual replications, though they would not label them as replications themselves. For this type of researcher, scientific advancement lies primarily in the generation of new ideas, new contexts, and novel applications. Replication studies, especially direct replications, are therefore referred to as unoriginal and of limited value.

*‘My main problem, in [my field of expertise], with many of the submissions I receive is that they tend to be rather trivial replications. People in [another country] are doing what we did ten years ago, for example. I’m not really interested in this kind of replication-type work; I’m more interested in novel ideas.’ (P7)*

However, within this view conceptual replications are deemed relevant, particularly when such studies involve applying existing theories to new contexts or populations. But there tends to be a resistance for the label of replication for this work. According to this view and in line with the views of the *skeptic*, conceptual replications are not replications per se, but rather new studies that happen to be informed by prior findings. The interest of the *innovator* lies in the expansion of knowledge, not in its verification.

The ideas of the *innovator* are linked to the view of not being concerned with whether a study confirms or rejects a prior result; instead, they seem to prefer information on the circumstances under which theories hold, i.e. what new

insights a study offers. If within this view someone were to analyze replication-like studies, they would prefer to trust in conventional statistical metrics, such as the  $p$ -value, to make the dichotomous decision on whether an effect is present. However, within this view researchers would prefer to see studies that are a continuation of broader scientific progress rather than a replication of a single study. During the interview the researcher asked the respondents to indicate how they would decide on corroboration of a replication if they ever would come across one, the respondent who fitted the characteristics of an *innovater* replied:

*‘We are still relying on p-value testing, so the results should be statistically significant. I’m not particularly interested in effect sizes or similar metrics, as they often cannot be compared, since we know that results tend to vary across different populations. Essentially, what matters is being able to draw the same conclusions; if that is the case, then it would be considered a corroboration.’*

(P7)

## 2.4 Discussion

With this study, we gained insight into the different experiences and opinions of researchers with respect to replication research. This study captures attitudes toward replication at a specific moment in time. It does not aim to assess how such attitudes evolve within individuals or across generations. Rather, it provides a snapshot of the field during a particularly noteworthy period. Following the initial shock of the presumed replication crisis around 2012, the scientific community began to respond and engage in critical self-reflection. We identified four different types of researchers based on their attitude towards replications. The insights into the four identified types offer an explanation for the ways in which replications occur in the literature.

The expressions of the interviewees connect to multiple themes identified in the existing literature. The respondents regularly discussed the rationale for conducting replication studies, although their views varied considerably. At one end of the spectrum, *collectivists* and *precisionists* argued that all studies are worth replicating and reporting, in line with, e.g., Crandall and Sherman (2016) or Lynch et al. (2015). At the other extreme, *innovators* regarded replication as having little or no value, as was previously stated by e.g. Feest (2019). The perspective of the *skeptic* emphasized the importance of replication studies in

the context of new insights, a view that aligns with the arguments put forward by scholars such as Finkel et al. (2017) and Field et al. (2019).

In addition, the respondents extensively reflected on how replications should be conducted. *Collectivists* expressed a preference for conceptual replications, which should ideally be carried out in collaboration with the original authors. This orientation appears to be driven by a search for what might be considered the “true answer,” a motivation also highlighted in earlier work (e.g., Crandall & Sherman 2016; LeBel & Peters, 2011). *Precisionists* and *skeptics* favored direct replications and preferred not to involve the original authors or only to verify if the design or set-up is similar to the original study. For both types, the primary value of replication lies in assessing whether the original findings were the result of chance or constitute a reliable reflection of an underlying effect, consistent with the perspectives articulated by, for example, Schmidt (2009) or Stroebe and Strack (2014).

The differing views were also reflected in the approaches of the respondents to evaluating corroboration. The proposed criteria ranged from the use of null hypothesis significance testing using  $p$ -values, as mostly mentioned by the *skeptic* and the *precisionist*, to more qualitative assessments of whether the replication results align with those of the original study, which is the preferred approach of *collectivists*. The *innovators* seem to be hard to place on this spectrum since they prefer to omit a decision about corroboration, but they indicate to use  $p$ -values or equivalents to determine significance. This range of analysis methods applicable to different goals has previously been discussed by Anderson and Maxwell (2016). In their paper Anderson and Maxwell explain the links per analysis method with different goals of replication, which in our typology is reflected by the attitude towards replication and the perceived primary goal of replication for each of the four types.

In the results, the career stage appears to be an interesting characteristic to explore further. In our sample, both the *constructivist* and the *precisionist*—each characterized by researchers in the early stages of their careers, i.e., at the level of assistant professor, post-doc, or PhD-candidates—expressed relatively positive attitudes toward replication. By contrast, the *innovator* and the *skeptic* were represented by researchers in later career stages, i.e., associate and full professors. However, the sample size in this study was small, with only a few researchers per type. Still, based on the available data, there seems to be a relation between the attitude on replication and career stage.

The small number of respondents has two consequences for this study. First, as mentioned before, although our sample suggests an association between career

stage and researcher type, this and other relationships cannot be established with certainty. The sample size is insufficient to conclude that these patterns would persist in a larger population. Second, the rigor of the four types is uncertain. In a larger sample, the typology might appear slightly different, with a different number of types and alternations in the definition of some of the types. Although the initial phase of analysis reached saturation, this cannot be assumed for the subsequent analytic stages.

Nevertheless, this study has documented some insight into the perspectives of psychology researchers during an ongoing period of change and debate within the disciplines of the social and behavioral sciences. Based on the empirical data obtained from researchers in the field, we created a typology which reflects the methodological literature discussed in the introduction of this paper offering a start to understand the views of groups of researchers.

## Acknowledgments

We are grateful to Herbert Hoijtink, Muriel Hagedaars, Fayette Klaassen, Henk Aarts, and Lakshmi Balachandran Nair for being involved in initiating this endeavor with Hidde Leplaa. They shaped the design of this study. We thank the people involved in the data collection: Herbert Hoijtink, Mariëlle Zondervan-Zwijnenburg, Lakshmi Balachandran Nair, Sjoerd Glaser, Karlijn Soppe, Vera Toepoel, and Duco Veen. We used ChatGPT for textual improvements.

## Bibliography

- Anderson, S., & Maxwell, S. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21, 1–12. <https://doi.org/10.1037/met0000051>
- Bettis, R. A., Helfat, C. E., & Shaver, J. M. (2016). The necessity, logic, and forms of replication. *Strategic Management Journal*, 37, 2193–2203. <https://doi.org/10.1002/smj.2580>
- Brandt, M., IJzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R., Grange, J., Perugini, N., Spies, J., & Van 't Veer, A. (2015). 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>
- Cesario, J. (2014). Priming, replication, and the hardest science [PMID: 26173239]. *Perspectives on Psychological Science*, 9(1), 40–48. <https://doi.org/10.1177/1745691613513470>
- Cohen, J. (1994). The earth is round (p i. 05). *American Psychologist*, 49(23), 997–1003. <https://doi.org/10.4324/9781315629049-14>
- Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress [Rigorous and Replicable Methods in Social Psychology]. *Journal of Experimental Social Psychology*, 66, 93–99. <https://doi.org/10.1016/j.jesp.2015.10.002>
- Etz, A., & Vandekerckhove, J. (2016). A bayesian perspective on the reproducibility project: Psychology. *PLoS one*, 11(2). <https://doi.org/10.1371/journal.pone.0149794>
- Feest, U. (2019). Why replication is overrated. *Philosophy of Science*, 86(5), 895–905. <https://doi.org/10.1086/705451>
- Field, S., Hoekstra, R., Brinkmann, L., & Van Ravenzwaaij, D. (2019). When and why to replicate: As easy as 1, 2, 3? *Collabra: Psychology*, 5(46). <https://doi.org/10.1525/collabra.218>

- Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology*, 113(2), 244–253. <https://doi.org/10.1037/pspi0000075>
- Glaser, B. (1978). *Theoretical sensitivity*. Mill Valley, CA: The Sociology Press.
- Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Psychology*, 66. <https://doi.org/10.1016/j.jesp.2015.09.009>
- Ioannidis, J. P. (2018). Why most published research findings are false. *Getting to Good: Research Integrity in the Biomedical Sciences*, 2(8), 2–8. <https://doi.org/10.1371/journal.pmed.0020124>
- Kolb, S. (2012). Grounded theory and the constant comparative method: Valid research strategies for educators. *Journal of Emerging Trends in Educational Research and Policy Studies*, 3, 83–86. <https://doi.org/10.10520/EJC135409>
- LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, 15(4), 371–379. <https://doi.org/10.1037/a0025172>
- Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. Sage. [https://doi.org/10.1016/0147-1767\(85\)90062-8](https://doi.org/10.1016/0147-1767(85)90062-8)
- Lynch, J. G., Bradlow, E. T., Huber, J. C., & Lehmann, D. R. (2015). Reflections on the replication corner: In praise of conceptual replications. *International Journal of Research in Marketing*, 32(4), 333–342. <https://doi.org/10.1016/j.ijresmar.2015.09.006>
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis?: What does 'failure to replicate' really mean? *American Psychologist*, 70(6), 487–498. <https://doi.org/10.1037/a0039400>
- McMullin, C. (2023). Transcription and qualitative methods: Implications for third sector research. *Voluntas*, 34, 140–153. <https://doi.org/10.1007/s11266-021-00400-3>
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Pro-

- moting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>
- Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, 18, 1–8. <https://doi.org/10.1371/journal.pbio.3000691>
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6), 615–631. <https://doi.org/10.1177/1745691612459058>
- Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6), 657–660. <https://doi.org/10.1177/1745691612462588>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, 6251. <https://doi.org/10.1126/science.aac4716>
- Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? three arguments examined [PMID: 26168109]. *Perspectives on Psychological Science*, 7(6), 531–536. <https://doi.org/10.1177/1745691612463401>
- Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528–530. <https://doi.org/10.1177/1745691612465253>
- QSR International Pty Ltd. (2020). Nvivo [Available from <https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home>].
- Rosenthal, R. (1979). The "file drawer problem" and tolerance for null effects. *Psychological Bulletin*, 86, 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Schmidt, S. (2009). Shall we really do it again? the powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2), 90–100. <https://doi.org/10.1037/a0015108>
- Shrout, P., & Rodgers, J. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, 69, 487–510. <https://doi.org/10.1146/annurev-psych-122216-011845>
- 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. <https://doi.org/10.1177/1745691612460687>

- Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication [PMID: 26173241]. *Perspectives on Psychological Science*, 9(1), 59–71. <https://doi.org/10.1177/1745691613514450>
- Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, 15(3), 1–18. <https://doi.org/10.1371/journal.pbio.2000797>
- Tryon, W. (1992). Thinking clearly about psychology: Volume 1: Matters of public interest: Essays in honor of paul e. meehl by d. cicchetti and w.m. grove (eds.). minneapolis: University of minnesota press, 1991. 277 pp. \$59.95. *Clinical Psychology Review*, 12(7), 787–788. [https://doi.org/10.1016/0272-7358\(92\)90024-3](https://doi.org/10.1016/0272-7358(92)90024-3)
- 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*, Volume 7 - 2016. <https://doi.org/10.3389/fpsyg.2016.01832>
- Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41, e120. <https://doi.org/10.1017/S0140525X17001972>

Appendix 2.A: Topic list Replication in Psychology

**Appendix 2.A: Topic list Replication in Psychology** The topic list used during the semi-structured interviews.

**Table 2.3:** The topic list

| Topics with subtopics | Check (Eliciting) questions                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Start                 | As introduced in our e-mail we investigate “What is Replication” by interviewing psychologists and methodologists from our faculty.                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Consent form          | For starters, did you read, agree with, and sign the consent form? What follows now are four main questions, some of which may have follow up questions.                                                                                                                                                                                                                                                                                                                                                                                                                       |
| You                   | Could you describe your field of research?                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| Field of research     | <p>Is there a role for replication studies in your field of research? <i>Follow up questions that can be asked if the topics below are not covered by the answer to the question above:</i></p> <ol style="list-style-type: none"><li>How did you form your opinion with respect to the role of replication studies in your field of research?<ol style="list-style-type: none"><li>discussions with colleagues;</li><li>literature;</li><li>newspaper articles?</li></ol></li><li>Do you think replication studies are actually executed in your field of research?</li></ol> |

Experiences  
with  
replication

Have you ever executed a replication study?

1. If yes, why? did you replicate your own work or that of others.
2. If no, why not? has your research ever been replicated by others?

Can you describe a core study from your own or someone else's research that you have replicated or would like to replicate? Why did you/would you like to replicate this research?

#### VIGNETTE A

What

What is a  
replication?  
Definition

What are the characteristics of a replication study?

*Follow up questions that can be asked if the topics below are not covered by the answer to the question above:*

1. Can replication studies be direct, conceptual, or of other types? Why or why not?
2. Should a replication study use the same measures as the original study?
3. Should a replication study use a sample from the same population as the original study?
4. Should a replication study have the same research design as the original study?
5. What is being replicated in a replication study?  
For example:
  - (a) a set of means;
  - (b) a p-value;
  - (c) a theory;
  - (d) order of a set of means.

How

What are the necessary steps to design and execute a replication study?

Set-up of a  
replication

*Follow up questions that can be asked if the topics  
below are not covered by the answer to the question  
above:*

Who  
replicates?

1. Who designs, executes, analyses, interprets, and reports a replication study?
2. What is the role of the authors of the original study?
3. What are requirements with respect to sample size, design, target population, and measurement instruments used?
4. If you executed a replication yourself, can you describe the process you experienced, from the idea to the analysis/report?

If you did not execute a replication, can you describe the steps you imagine you would go through?

VIGNETTE  
B

If you were to conduct a replication of the study mentioned in the vignette, how would you do it?

1. Would you do a direct/conceptual/other type of replication?
2. What should be replicated in this case?
3. Would you use:
  - (a) the same variables measures in the same manner as in the original study? (Why/why not?)
  - (b) a sample from the same population as the original study? (Why/why not?)
  - (c) the same research design as the original study? (Why/why not?)

If you were to conduct a replication of one of your own studies would you do something differently? (Why/why not?)

---

Outcome

When do you consider the outcomes of an original study to be replicated?

When is it replicated

*Follow up questions that can be asked if the topics below are not covered by the answer to the question above:*

1. Should, for example, means and effect sizes from original and replication studies be the same? OR, should a pattern in the means in the original and replication studies be the same?
2. How can one determine whether either of the above holds? What are the criteria?

VIGNETTE  
C

---

Ending

Is there anything important relevant to the topic of replication that we haven't discussed yet.

Thank you for your participation you can expect the transcript of this manuscript for your approval in a week. If you have any remarks or comments, sent them to us in a month please. If you think of anything else to add regarding the topics we discussed feel free to contact us within this week.

---

### **Vignette A**

Our motivation for this question are the outcomes of the “Reproducibility Project Psychology” by the “Open Science Collaboration” that showed that of 100 original studies, only the results of about 30 studies could be replicated.

Does this have implications for your field?

## Vignette B

Siegle and Schuler (2000) collected data from 391 gifted sixth through eighth graders from various socioeconomic regions of New England, and the Mid-Atlantic and Southern parts of the United States. In their study, Siegle and Schuler explored all differences among grade levels, birth order positions, and gender in academically talented middle grade students using a modified version of the FMPS. They also examined the appropriateness of using the Goals and Work Habits Survey (Schuler, 1994), a modified version of the FMPS (Frost Multidimensional Perfectionism Scale) with their sample. A between-subjects factorial MANOVA examined differences among grade level (grades 6, 7, 8), birth order (first or only, middle, youngest), and gender (male, female). The MANOVA and a Principal Components Analysis (PCA) found significant gender differences where females had higher organization than males, and males felt greater expectations from their parents than females. Birth-order differences were found between the first-born and youngest child, where first-born children felt higher levels of parental criticism and expectations. Gender and grade interacted significantly with females' concern over mistakes increasing as they grew older and males' concern over mistakes peaking in seventh grade. Parental criticism was highest for females in grades 6 and 8, where males had greatest concerns over parental criticism in seventh grade. Finally, Siegle and Schuler (2000) interpreted their results as an indication that their modified version of the FMPS could be used to measure perfectionism with academically talented students in grades 6 through 8.

From the article: Sondergeld, T. A., Schultz, R. A., & Glover, L. K. (2007). The need for research replication: An example from studies on perfectionism and gifted early adolescents. *Roeper Review*, 29(5), 19-25.

### Question

If you were to conduct a replication of the study mentioned in the vignette, how would you do it?

Vignette C

Instruction

We present you with the results of an original study and a direct replication of that study. We would like you to take a look at the graphs, and explain how you would decide if this was a successful replication. You may very well indicate that you need more information, please let us know what you need to determine the successfulness of the replication.

Figure 2.1: Results original study

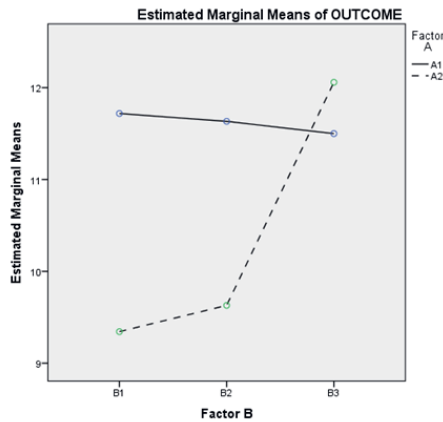
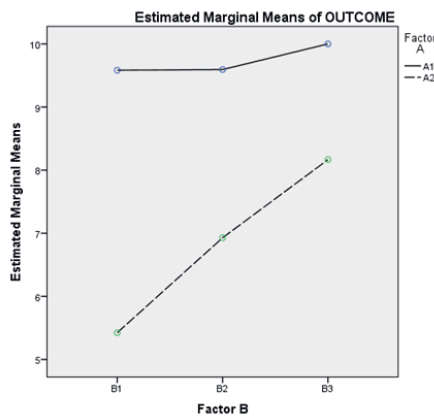


Figure 2.2: Results replication study



## CHAPTER 3

### Bayesian evaluation of replication studies

Leplaa, H.J., Rietbergen, C., & Hoijsink, H. (2024). Bayesian Evaluation of Replication Studies. *Meta-Psychology*, 8. DOI: <https://doi.org/10.15626/MP.2020.2554>

#### Abstract

In this paper a method is proposed to determine whether the result from an original study is corroborated in a replication study. The paper is illustrated using two replication studies and the corresponding original studies from the reproducibility project psychology by the Open Science Collaboration. This method emphasizes the need to determine what one wants to replicate from the original paper. This can be done by translating the research hypotheses formulated in the introduction into informative hypotheses, or, by translating the results into interval hypotheses. The Bayes factor will be used to determine whether the hypotheses resulting from the original study are corroborated by the replication study. Our method to assess the successfulness of replication will better fit the needs and desires of researchers in fields that use replication studies.

*Keywords:* Bayes factor, Informative Hypothesis, Replication Crisis, Replication Study

---

Author contribution: HL, CR, and HH were involved in the initial research design. HL wrote and revised the article in collaboration with HH, CR was actively involved in the review and edit of the first draft. HL conducted the simulation studies and the analyses.

### 3.1 Introduction

People will walk more slowly when primed with words related to being old was the result of a study by Bargh et al. (1996). However, the replication study by Doyen et al. (2012) failed to corroborate the results. This study became one of the examples of the existence of the so-called replication crisis in the behavioral sciences; results obtained in a study are often not corroborated by its replication. The question is now: what should we believe? How can we rely on results from scientific research in the behavioral sciences if they contradict each other so often? But first: do they actually contradict each other? We will focus on the question how it can be determined whether a replication study supports the earlier study. We will propose two methods to translate an original paper into a hypothesis: either based on the theory, or on the results. Subsequently, the data resulting from the replication study will be used to evaluate these hypotheses with the Bayes factor.

We will illustrate this paper with two studies from the Open Science Collaboration (OSC) Reproducibility Project Psychology (OSC, 2012, 2015). For each of 100 psychological experiments, an exact (Hüffmeier et al., 2016) replication study was executed by researchers from the OSC, using a procedure as similar as possible to the original study and indicating in detail any methodological differences between the original study and the replication study. Furthermore, the sample size for the replication study was determined based on a power analysis. The statistical analyses in the replication studies were the same ones as used in the original studies. The goal was to evaluate whether the results of the replication study corroborated the original study; which would mean it was a successful replication. In this project, only 36% to 47% of the original studies was successfully replicated. Since the OSC consists mainly of studies comparing group means, our focus will be on ANOVAs, although the ideas can easily be translated to other statistical models.

It is open for debate how many studies within the OSC (2012, 2015) were replicated. This is due to the fact that different methods were used to determine the success of the replications. Throughout the literature, different kinds of methods are used to evaluate replication studies. Anderson and Maxwell (2016) identify six different goals when evaluating successfulness of replication studies, and present for each goal a suitable method to analyse it: *significance-based*, *null effect inferring*, *effect size estimating*, *meta-analysis*, *assessing inconsistency*, and *assessing consistency*. We will discuss these methods, along with some of the critiques. Both the third, *effect size estimation*, and fourth, *meta-analysis*,

do not focus on testing whether a replication study corroborates the results of the original study. Therefore we will not discuss these methods. Then, we will introduce our methods as an additional and innovative option because in contrast to the other approaches we formulate the replication hypotheses explicitly based on the theory or results presented in the original paper.

The *significance-based method* was done by testing whether the effect observed in both the original and the replication study was significant. It is also referred to as *vote counting* (Anderson & Maxwell, 2016; Simonsohn, 2015; Zondervan-Zwijnenburg et al., 2019). If the statistical tests provide the same answer (e.g. both significant or both non-significant), and the direction of the effect is the same, the study is considered to be replicated. The problem with this approach is that while an original study with a  $p$ -value of .04 is not considered to be replicated if the replication study finds a  $p$ -value of .06, the same original study is replicated when the replication study finds a  $p$ -value of  $< .001$ . Interestingly, the difference between a  $p$ -value of .04 and .06 is not necessarily significant itself Gelman and Stern (2006). For further research in line with the significance-based method we refer to Etz and Vandekerckhove (2016), Ly et al., (2018), Marsman et al., (2017), Simonsohn, (2015), or Verhagen and Wagenmakers, (2014).

The second method discussed in Anderson and Maxwell (2016) was *null effect inferring*. In this method, usually called equivalence testing, the null hypothesis that there is a relevant effect is tested against the alternative hypothesis, that there is not a relevant effect. This is done by establishing a "region of indifference" (Anderson & Maxwell, 2016). This region identifies which results are essentially zero, assuming that finding an effect of exactly zero is nearly impossible. If then the 90%-confidence interval (90%-CI) around the effect found in the replication falls completely in this region of indifference, the hypothesis that there is a relevant effect has to be rejected in favor of the hypothesis that there is not a relevant effect. If it (partially) lies outside this region, the hypothesis that there is relevant effect can not be rejected. An issue with this method is that it is difficult to determine the width of the region of indifference. The interested reader is referred to Simonsohn (2015) who specifies such an interval using a Cohen's  $d$  of .1, however, without providing an argument for this specific interval. Furthermore, the results of the original study are disregarded when deciding to accept or reject the null hypothesis. That is, the region of indifference is determined without taking into account the data and results from the original study, in the next method the original study does have a role when determining the hypotheses of interest.

*Assessing inconsistency* is an answer to the critique that the difference between significant and non-significant is often not significant itself (Gelman & Stern, 2006). It is a test of the heterogeneity of the effect sizes resulting from the original and replication study. It is performed by calculating a confidence interval around the difference in effect sizes. It is a sophisticated method to assess the similarity between the original and replication study. Since only one effect size per study can be included in the analysis at once the method of *Assessing inconsistency* only properly works for *t*-tests, which makes it somewhat limited. Our method will extend on this method but will be suitable for ANOVAs as well. Another application of this method can be found in Patil et al. (2016), though their method is limited to a confidence interval for the original study. Furthermore, as Morey and Lakens (2016) also highlighted, their method requires well-powered original and replication studies. The latter is not unique to Patil et al.'s approach, and will further be discussed later in this paper.

Last Anderson and Maxwell (2016) identify the method called *assessing consistency*. This method combines the methods two (equivalence testing) and five (assessing inconsistency): they test if the CI around the difference in the effect sizes falls within a subjectively determined region of indifference. As with the second method: the width of the interval is subjectively determined. But, as Anderson and Maxwell (2016) mention themselves: only with very large sample sizes the CI will be small enough to reasonably expect it to fit within the region of indifference. In all applications without very large sample sizes, this method will never confirm that the effect sizes are equal.

Three issues apply to both the first two and the last two methods identified by Anderson and Maxwell (2016). First, all methods result in a dichotomous interpretation of the test results: either the results from the original study are replicated or they are not. However, although this is often ignored, there are error probabilities associated with these decisions: if the null hypothesis is rejected there is usually a Type I error of 5% probability that this is incorrect, and if the null hypothesis is not rejected, there is an unknown (because the population effect size is not known) Type II error probability that the null hypothesis is incorrectly not rejected. Second, the methods test just one hypothesis at a time. As we will show in this paper, sometimes researchers want to test multiple competing hypotheses. This can be the case in, for example, a two-way ANOVA, where the goal may be to determine if none, one, or both possible main effects are present (as is the case in the study by Janssen et al., 2008). Third and last, all methods test a (interval) hypothesis against the corresponding alternative hypothesis, as will be shown in this paper our approach also allows for a class

of more flexible hypotheses that are called informative hypotheses (Hojtink, 2012).

The approaches proposed in this paper is a new and innovative addition to the approaches identified by Anderson and Maxwell (2016). First, we use the Bayes factor (BF) which replaces dichotomous decisions by a quantification of the support in the data for each of the hypotheses involved. Use of support as quantified by the Bayes factor is elaborated by Morey et al. (2016). It is an alternative for Popperian falsification (Popper, 1963) called logical positivism. The interested reader is referred to Hawthorne (2021) for an elaboration of logical positivism, that is, the philosophical foundation of using the BF as a measure of support for the hypotheses of interest. Second, we allow for the evaluation of multiple hypotheses at once, with the additional benefit that the methods can be applied to ANOVAs and *t*-tests while the other approaches are mostly tailored to the *t*-test. Third, we force the replication researcher to make an informed decision about the formulation of the replication hypotheses.

The BF (Hojtink, Gu, et al., 2019b; Kass & Raftery, 1995) was already proposed by Anderson and Maxwell (2016) to evaluate the results of replication studies. As will be elaborated later, the Bayes factor measures the relative support in the data for two or more competing hypotheses. The BF has been used previously to evaluate replication studies. Some of these studies used the ANOVA or *t*-test null hypotheses (Etz & Vandekerckhove, 2016; Harms, 2018; Ly et al., 2018; Marsman et al., 2017) versus a one- or two-tailed alternative hypothesis. Other studies using the BF only focused on *t*-tests (Field et al., 2019; Marsman et al., 2017; Van Aert & Van Assen, 2017).

Cho and Abe (2013) addressed the fact that researchers do not always test the correct hypothesis. Often, this involves testing the null hypothesis, which was previously criticized by Gigerenzer (2004). We argue that oftentimes retesting the null hypothesis is not in line with the goals of the replication researchers, who want to determine if the replication study corroborates the results from the earlier study. To test more relevant hypotheses, we allow the replication researchers to evaluate replication hypotheses based on either the introduction of the original paper (subsequently called the "theory based method") or the results of the original paper (subsequently called the "results based method"). This is not unique to our approach, other examples can be found in Simonsohn (2015), who use the results of the original study to determine the effect size to be evaluated in a replication study. Another example, although not specific to the evaluation of replication studies, is replacing the null hypothesis with an interval hypothesis, as is done in equivalence testing (Lakens et al., 2013).

In our methods we will evaluate other kinds of hypotheses than used in these articles. The theory based method translates the introduction section from an original paper into an informative hypothesis or hypotheses (Hoijsink, 2012; e.g.  $H_t : \mu_3 < \mu_2 < \mu_1$ ), whereas the results based method uses the results section to formulate interval hypotheses (as in, e.g. Lakens et al., 2013, Zondervan-Zwijnenburg et al., 2019; e.g.  $H_r : 5.28 < \mu_1 < 5.94 \ \& \ 4.90 < \mu_2 < 5.56$ ). After data is collected the BF is calculated to determine the relative support in the data for the replication hypothesis (the informative or interval hypothesis) versus its complement. This renders an innovative method that uses informative hypotheses and the BF to evaluate replication studies. Our approach explicitly addresses the fact that researchers do not always test the correct hypothesis (Cho & Abe, 2013) and therefore is a remedy against the "null-ritual", which was previously criticized by Gigerenzer (2004). Our paper is also in line with Dienes (2014) who discusses Bayesian interpretations of non-significant results (however, not in the context of replication studies). To address a similar issue, we step away from using the null-hypothesis, instead use informative and interval hypotheses, and put this to work in the context of replication studies.

This paper is structured as follows: our method is illustrated using one of the replication studies from the OSC (2012; 2015). This study will be introduced in the next section. Following this introduction, we will explain and demonstrate our methods, where an original paper is translated to hypotheses, from start to finish. Subsequently, a second example employing more than one hypothesis of interest will be presented. This paper is concluded with a discussion in which we will reflect on the possibility to synthesize information from different studies, similar to methods three and four of Anderson and Maxwell (2016).

## 3.2 The running example

The running example used in this paper is a study on the perceptions of closeness toward one's family members and hometown (Williams & Bargh, 2008). In Study 4 of this paper, participants were instructed to mark off two points on a Cartesian coordinate plan: these could be close to the origin (Group 1 - Closeness), some distance from the origin (Group 2 - Intermediate), or far from the origin (Group 3 - Distance). Participants then rated the strength of their bonds to their siblings, their parents, and their hometown. All questions were answered on a scale ranging from 1 (*not at all strong*) to 7 (*extremely strong*) and averaged to create a total score. The hypothesis was that marking points more distant from the origin would result in perceiving the closeness toward one's

family members and hometown as less strong. The results from this study are summarized in Table 3.1, which contains more information and will be referred to throughout this paper. Williams and Bargh (2008) concluded that their expectations were matched by their results. This can be seen from the  $p$ -value of .01, which indicates that the three means are not equal, and an ordering of the means as hypothesized.

Joy-Gaba et al. (2012) executed a direct replication of the study by Williams and Bargh (2008). Their procedures were as similar as possible, using the same materials in the replication study as were used in the original study. The results from the replication study, also presented in Table 3.1, show no significant differences between the groups. Using the procedure of *vote counting* it can be concluded that the results were not successfully replicated: the  $p$ -value is not below the threshold of .05.

### 3.3 What should be replicated

As elaborated in the introduction, our method to evaluate replication studies determines the support in the replication data for one or more hypotheses resulting from the original study. Two methods to obtain these hypotheses will be presented in this section. In the first, the hypotheses are extracted from the theoretical framework as presented in an introduction of the original paper. In the second, hypotheses are formulated based on the information in the results section of the original paper. An overview of the steps is given in Table 3.2.

#### 3.3.1 Method 1 - Theory based method

To construct hypotheses from the introduction of the original study a six step procedure will be used that is based on qualitative research methods (e.g. Glaser, 1978; Charmaz, 2006).

##### Step 1.1 - Reading the introduction

Read the introduction section of the original paper. The introduction typically contains the theoretical background for the study, and results in the formulation of theory-based hypotheses that will be tested in the study. There is no output from this step.

**Table 3.1:** Descriptive statistics and results of the original study of Williams and Bargh (2008) and its replication by Joy-Gaba et al., 2012.

|                                                      | Original study by Williams<br>& Bargh (2008) |              |              | Replication study by Joy-Gaba<br>et al. (2012) |              |              |
|------------------------------------------------------|----------------------------------------------|--------------|--------------|------------------------------------------------|--------------|--------------|
|                                                      | Group 1                                      | Group 2      | Group 3      | Group 1                                        | Group 2      | Group 3      |
|                                                      | Closeness                                    | Intermediate | Distance     | Closeness                                      | Intermediate | Distance     |
| $M$ ( $SD$ )                                         | 5.61 (0.90)                                  | 5.23 (0.90)  | 4.86 (0.90)  | 5.44 (0.83)                                    | 5.31 (1.07)  | 5.31 (1.15)  |
| 95% CI                                               | (5.28, 5.94)                                 | (4.90, 5.56) | (4.53, 5.19) | (5.19, 5.69)                                   | (4.99, 5.63) | (4.94, 5.68) |
| $n$                                                  | 28                                           | 28           | 28           | 44                                             | 43           | 38           |
| $M_{1-2}$ (95% CI)                                   | 0.38 (-0.09, 0.85)                           |              |              | 0.13 (-0.27, 0.53)                             |              |              |
| $M_{2-3}$ (95% CI)                                   | 0.37 (-0.10, 0.84)                           |              |              | 0.00 (-0.48, 0.48)                             |              |              |
| $H_0 : \mu_1 = \mu_2 = \mu_3$                        | $p = .01, \eta^2 = .11$                      |              |              | $p = .79, \eta^2 = .004$                       |              |              |
| $H_{ts}^* : \mu_3 < \mu_2 < \mu_1$                   |                                              |              |              | $BF_{ts^c} = 2.10$                             |              |              |
| $H_{t_{min1}} : \mu_3 + 0.21 < \mu_2 + 0.10 < \mu_1$ |                                              |              |              | $BF_{t_{min1}^c} = 0.57$                       |              |              |
| $H_{t_{min2}} : \mu_3 + 0.41 < \mu_2 + 0.19 < \mu_1$ |                                              |              |              | $BF_{t_{min2}^c} = 0.11$                       |              |              |
| $H_{r_M} : 5.28 < \mu_1 < 5.94 \ \&$                 |                                              |              |              | $BF_{r_M^c} = 1.92$                            |              |              |
| $4.90 < \mu_2 < 5.56 \ \& \ 4.53 < \mu_3 < 5.19$     |                                              |              |              |                                                |              |              |
| $H_{r_D} : -0.09 < \mu_1 - \mu_2 < 0.85 \ \&$        |                                              |              |              | $BF_{r_D^c} = 4.44$                            |              |              |
| $-0.10 < \mu_2 - \mu_3 < 0.84$                       |                                              |              |              |                                                |              |              |

Note that, Williams and Bargh (2008) do not present their three means, standard deviations and sample sizes. We were unable to retrieve the original data, therefore numbers presented in the table were reconstructed using the information presented in the paper assuming equal within groups variances and assuming that the mean of the second group was located on an equal distance of the other two means. This is very likely not in line with the original data, but allows us to illustrate our approach using a real original study.

\*= In the introduction section Williams & Bargh only explicitly discuss the groups Closeness and Distance, but in line with the theory underlying distance ques, we assumed that the effect of "Intermediate" is between that of "Distance" and "Closeness".

### Step 1.2 - Coding of statements in the introduction

Those statements are coded that specify theories or expectations with respect to the outcomes of the study. Coding is a qualitative (Glaser, 1978; Charmaz, 2006; Boeije, 2010 p.79) manner of data analysis. Typically, coding consists of two activities: segmenting and labeling (textual) data. For our purposes, the labeling part is not that relevant, because the only goal is to localize statements that can be labeled "theory or expectation with respect to the outcomes of the study", that is, for each statement selected the label is the same. The statements of interest are often indicated with words such as "hypothes\*", "expect\*", and "argue\*" <sup>1</sup>.

<sup>1</sup>\*= different extensions are possible, e.g. hypothesis, hypothesize, hypotheses

**Table 3.2:** Overview and short explanation of the steps.

| Theory based method                                                                          | Results based method                                                         |
|----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------|
| <b>1.1 - Reading the introduction</b>                                                        | <b>2.1 - Reading the results section</b>                                     |
| Simply read through the section to know what is being studied                                | Simply read through the section to know what is being studied                |
| <b>1.2 - Coding statements in the introduction</b>                                           | <b>2.2 - Extracting the results</b>                                          |
| Indicate the prior expectations of the researchers                                           | Extract the results and indicate what the results of the conducted tests are |
| <b>1.3 - Reading the methods and the results section</b>                                     | <b>2.3 - Translating results to interval hypotheses</b>                      |
| Simply read through the section to know what tests were conducted                            | Calculate the desired interval hypotheses based on the results               |
| <b>1.4 - Coding of statements in the methods and the results section</b>                     |                                                                              |
| Select the statements that report the results of the tests                                   |                                                                              |
| <b>1.5 - Selecting the statements from the introduction that are actually tested</b>         |                                                                              |
| Select the statements from Step 1.2 that are tested with the statements selected in Step 1.4 |                                                                              |
| <b>1.6 - Translating statement to informative hypotheses</b>                                 |                                                                              |
| Formulate the tested statements from the introduction section as testable hypotheses         |                                                                              |

This step gives as output one or more statements. In the running example two statements were coded:

1. "Accordingly, we argue that a primitive understanding of physical distance is the foundation for the later-developed concept of psychological distance, given humans' pervasive tendency to conceptualize the mental world by analogy to the physical world" (Williams & Bargh, 2008, p.2)
2. "We hypothesized that participants primed with distance would report weaker attachments to their family members and hometown, compared with participants primed with closeness" (Williams & Bargh, 2008, p.6)

### **Step 1.3 - Reading the methods and the results section**

Read the results section of the original paper. This section elaborates which analyses are to be executed. No output results from this step.

### **Step 1.4 - Coding of statements in the methods and the results section**

Statements are coded that indicate what is actually tested. In theory, this can be statements in the methods section explaining which tests will be conducted to test the hypotheses of interest. However, after checking 25 papers from the OSC reproducibility project psychology to find examples of ways to present this information in the methods section, we had to conclude that this information is never, or barely ever, presented in the methods section. Therefore only the results section is used for the following steps, since here the relevant information is always present. It remains necessary to read the methods section, to understand how the variables used in a study are operationalized.

In the results section the outcomes of the tested hypotheses are presented. However, note that, these will only be coded according to which hypothesis has been evaluated with which test, not to clarify the results of that test. The statements of interest may be indicated with words such as: "we have tested (...)", "we used/conducted (...) analysis/test (...)", and "to test the hypothesis (...)". Also, sentences indicating particular tests are of interest, e.g. *t*-test, ANOVA, and regression. For the running example the following statements were coded in the results section:

1. "Next, we conducted a planned contrast analysis using weights of -1, 0, and 1 for a linear contrast, and -1, +2, -1 for a quadratic contrast, for the Closeness, Intermediate, and Distance conditions, respectively. These contrast weights allowed us to test the specific hypothesis that participants in the Distance condition enjoyed the excerpt more than those in the Intermediate condition, who in turn enjoyed it more than those in the Closeness condition." (Williams & Bargh, 2008, p.4)
2. "An ANOVA revealed that the three spatial-prime groups differed significantly in the reported strength of their bonds to their family and hometown, (...)" (Williams & Bargh, 2008, p.7)

This step results in one or more coded statements.

### Step 1.5 - Selecting the statements from the introduction that are actually tested

The statements coded in Step 1.4 can be used to verify which statements coded in Step 1.2 are actually tested. For the running example, as can be seen from the second statement coded in Step 1.4, only the second statement coded in Step 1.2 ("We hypothesized that participants primed with distance would report weaker attachments to their family members and hometown, compared with participants primed with closeness") was tested. The result of Step 1.5 is a number of statements.

### Step 1.6 - Translating statements to informative hypotheses

In Step 1.6 informative hypotheses are formulated. In the theory based method only ordered hypotheses are formulated, of which two forms are distinguished: simple ordering and minimal difference hypotheses on (combinations of) parameters.

Informative hypotheses are a formal representation of the expectations a researcher has with respect to the relations between the means in an ANOVA (Klugkist et al., 2005; Hoijtink, 2012, p. 50-51). Using inequality constraints, order relations between (combinations of) means ( $\mu$ ) can be specified. Three examples are:

$$H_1 : \mu_1 < \mu_2 < \mu_3,$$

$$H_2 : \mu_1 + 0.2 < \mu_2 \text{ \& } \mu_1 + 0.5 < \mu_3,$$

and

$$H_3 : \mu_{1A} - \mu_{1B} > \mu_{2A} - \mu_{2B}.$$

$H_1$  shows a specific ordering of three means, where the mean of Group 1 is smaller than that of Group 2, which in turn is smaller than that of Group 3.  $H_2$  is also an ordering of three means, but now the minimal difference is specified. The mean of Group 1 is at least 0.2 points lower than the mean of Group 2, and 0.5 points lower than the mean of Group 3.  $H_3$  shows a two way analysis of variance, with two factors with levels 1/2 and A/B. Here, it is stated that the difference between the means of Levels A and B is greater when located at Level 1, than when located at Level 2. This is the start of a hypothesis that specifies an interaction effect. To further specify the interaction effect, elements like  $\mu_{1A} > \mu_{1B}$  and  $\mu_{1A} > \mu_{2A}$  have to be added.

Next to inequality constraints, one could also use equality constraints. Two examples are:

$$H_4 : \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B},$$

and

$$H_5 : \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B} \ \& \ \mu_{1A} + \mu_{2A} > \mu_{1B} + \mu_{2B}.$$

$H_4$  indicates the absence of a main effect of the factor with levels 1/2. No specification has been given in  $H_4$  regarding a main effect of the factor with levels A/B or an interaction effect of those two factors. Next,  $H_5$  is an extension of  $H_4$ . Next to the specified absence of a main effect of the factor with levels 1/2, a main effect of the factor with levels A/B is specified here.  $H_5$  hypothesizes that there is main effect, with the subjects located at level A score on average higher than subjects located at level B.

To formulate the informative hypothesis based on the theory section of the paper, first the dependent variable needs to be determined. Second, it needs to be determined which groups are distinguished and the theory needs to be translated into inequality and or equality restrictions on the group means. Applied to the running example this renders: the variables in the statement ("We hypothesized that participants primed with distance would report weaker attachments to their family members and hometown, compared with participants primed with closeness"; Williams & Bargh, 2008, p.6) are 1) attachment to family members and hometown, and 2) the prime. The first variable, attachment, is measured on a 7-point interval scale. For the prime variable three groups are distinguished in this study; Distance, Intermediate and Closeness. The groups are clearly ordered in this statement: a group primed with more closeness scores is expected to score on average higher on the dependent variable, strength of the attachment, than the group primed with more distance. Because there is no mention of the size of the difference, an informative hypothesis with an inequality constraint will be formulated:

$$H_{t_s} : \mu_{Distance} < \mu_{Intermediate} < \mu_{Closeness},$$

where, the subscript  $t$  denotes that the hypothesis is based on the literature review in the introduction of the paper, and the subsubscript  $s$  indicates a simple ordering hypothesis is formulated. We would like to add another variant of these types of hypotheses. Typically, when discussing differences between two means, people refer to a relevant difference. The hypothesis  $H_{t_s}$  leaves

room for irrelevant differences, e.g. the mean of the Distance group is 0.01 lower than the mean of the Closeness group. In the second hypothesis based on the introduction section we assume at least a small effect, that is a minimal difference of 0.2 standard deviation (*SD*; Cohen, 1988). There are two ways to interpret this: either 0.2 *SD* difference between adjacent groups or between the groups with the most extreme means. To show the effect of either decision, two hypotheses are formulated:

$$H_{t_{min1}} : \mu_{Distance} + 0.21 < \mu_{Intermediate} + 0.10 < \mu_{Closeness}$$

and

$$H_{t_{min2}} : \mu_{Distance} + 0.41 < \mu_{Intermediate} + 0.19 < \mu_{Closeness}$$

where the subsubscript *min* indicates that a minimal difference hypothesis is formulated with at least Cohen's *D* equals 0.2 between adjacent groups. The score of 0.10 equals 0.1 times the pooled within group standard deviation of groups Intermediate and Closeness in the replication data. The score of 0.19 equals 0.2 times the pooled within group standard deviation of groups Intermediate and Closeness in the replication data. The scores of 0.21 and 0.41 equal 0.1 and 0.2 times the pooled within group standard deviation of groups Distance and Intermediate on top of the difference between the groups Intermediate and Closeness: i.e. 0.21 is based on 0.0956 (0.1 times pooled within group *SD* for Intermediate and Closeness) + 0.1108 (0.1 times pooled within group *SD* for Distance and Intermediate). The result of this step is an informative hypothesis that can later be evaluated using the data resulting from the replication study. Before we continue with the analysis, we will discuss the results based method.

### 3.3.2 Method 2 - Results based method

To construct hypotheses based on the outcomes of the original study a three step procedure will be used.

#### Step 2.1 - Reading the results section

Read the results section of the original paper, which reports the outcomes of the conducted analyses. It might also be useful to read the methods section to get a general idea of how the study was executed. There is no output from this step.

**Step 2.2 - Extracting the results**

Extract the results from the original paper. The relevant results for this method are the mean and its standard error for each group. If the standard error is not reported it can be computed using the standard deviation and sample size for each of the groups. Those statistics are often included in a table with descriptive statistics in the results section. In the running example not all the needed statistics were reported. The available statements that were used to determine the results reported in Table 3.1 are:

1. "An ANOVA revealed that the three spatial-prime groups differed significantly in the reported strength of their bonds to their family and hometown,  $F(2, 81) = 4.97$ ,  $p_{rep} = .95$ ,  $\eta^2 = .11$ ." Note that, this corresponds to the  $p$ -value of .01 as reported in Table 3.1.
2. "(...) the linear contrast showed that participants primed with distance reported weaker bonds to their family and hometown ( $M = 4.86$ ), compared with participants primed with closeness ( $M = 5.61$ ),  $t(81) = -2.86$ ,  $p_{rep} = .96$ ." Note that, this corresponds to the  $p$ -value of .01 as reported in Table 3.1.

**Step 2.3 - Translating results to interval hypotheses**

With interval hypotheses it can be tested whether one or more means are within or outside of a certain interval of values (Zondervan-Zwijnenburg et al., 2019). Examples of interval hypotheses are:

$$H_6 = 0.2 < \mu_1 < 0.4,$$

$$H_7 = -0.1 < \mu_1 < 0.7 \ \& \ 1.1 < \mu_2 < 2.4,$$

and

$$H_8 = -0.6 < \mu_1 - \mu_2 < 0.1 \ \& \ 0.7 < \mu_2 - \mu_3 < 1.2.$$

In  $H_6$  it is stated that the mean of Group 1 is within the range 0.2 to 0.4. Whereas  $H_7$  states that the mean of Group 1 is somewhere in the range -0.1 through 0.7, and the mean of Group 2 is in the interval 1.1 to 2.4. Lastly,  $H_8$  states that the difference between the means of Groups 1 and 2 is somewhere

in the range of -0.6 ( $\mu_2$  is considerably larger) to 0.1 ( $\mu_1$  is slightly larger), and the difference between the means of Groups 2 and 3 is in the range of 0.7 to 1.2 ( $\mu_2$  is larger). The end points of the interval should be chosen such that they represent the uncertainty in the estimates of the means obtained from the original study. As can be seen in Equations 3.1 and 3.2 the lower bound (LB) of the interval is the lower bound of the traditional 95% confidence interval for the mean or the difference between two means, and the upper bound (UB), Equations 3.3 and 3.4, is chosen analogously:

$$LB_{mean} = M - 1.96 * \frac{SD}{\sqrt{n}} \quad (3.1)$$

$$LB_{difference} = M_1 - M_2 - 1.96 * \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}} \quad (3.2)$$

$$UB_{mean} = M + 1.96 * \frac{SD}{\sqrt{n}} \quad (3.3)$$

$$UB_{difference} = M_1 - M_2 + 1.96 * \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}, \quad (3.4)$$

where all statistics are derived from the original study.

Due to publication bias (Rosenthal, 1979)  $M_1 - M_2$  is possibly an overestimation of the true effect. If this is indeed the case,  $M_1 - M_2$  obtained from the replication data will probably be (much) smaller and not contained in the intervals specified in the hypotheses. Therefore, the support in the replication data for the replication hypotheses will be small, which is, in this case, the desired result. If the interval in the replication hypothesis becomes very narrow, this would be the case when the sample size in the original study was large, two things can happen. Either the result from the replication study falls within the interval, which would lead to strong support for the replication hypothesis as is desired. Or, if bias was present in the original study, the results from the replication study falls outside of the interval which would lead to a strong rejection of the replication hypothesis, which is also desired.

To formulate an interval hypothesis, a similar procedure should be followed as for the theory based method. Start with determining the dependent variable. Then, determine which groups are distinguished. Finally, the confidence interval for each group can be determined. In the running example, three means were reported. For each of those means an interval has to be specified. Using the descriptive statistics presented in Table 3.1 and Equations 3.1 and 3.3 results in:

$$H_{rM} : 5.28 < \mu_{Closeness} < 5.94 \ \& \ 4.90 < \mu_{Intermediate} < 5.56 \ \& \ 4.53 < \mu_{Distance} < 5.19,$$

where the subscript  $r$  indicates the hypothesis is formulated based on the results, and the subsubscript  $M$  indicates it is an interval hypothesis based on the means. There is another possibility however. Instead of focusing on the means, one could also focus on the difference between means. Often, researchers are not necessarily interested in the means themselves, but rather on the respective position of one mean related to another. Therefore, interval hypotheses could also be formulated such that they specify in what region the difference between two means fall. For this study it results in the following hypothesis:

$$H_{rD} : -0.09 < \mu_{Closeness} - \mu_{Intermediate} < 0.85 \ \& \ -0.10 < \mu_{Intermediate} - \mu_{Distance} < 0.84,$$

where the subsubscript  $D$  indicates the interval hypothesis is based on the difference between the means. In hypothesis  $H_{rD}$  it is specified that the difference between the means of the groups Closeness and Intermediate should be between -0.09 (mean of the Intermediate group is somewhat higher) and 0.85 (mean of Closeness is considerably higher), and the difference between the means of the groups Intermediate and Distance is in the range of -0.10 to 0.84. This interval is based on the 95%-CI calculated around the difference of these means in the original study. Furthermore, the difference between the means of the Intermediate and Distance group is also specified in hypothesis  $H_{rD}$ .

At this point, the replication hypotheses are formulated in the form of informative hypotheses in the theory based method or in the form of interval hypotheses in the results based method. In the following section we discuss how these hypotheses can be evaluated.

### 3.3.3 An important property of the original study

An important note needs to be made. When using the original paper to formulate hypotheses, the sample size of that original study is of high influence on usefulness of the replication (Anderson et al., 2017; Anderson & Maxwell, 2016). Note that this only influences the potential usefulness of replications conducting according to the results based method. These problems are not present in the theory based method, as we will discuss below. We specify two situations that both seem problematic: a large and a small sample size in the original

study. This phenomenon is touched upon in the previous subsection (and is also explored in Simonsohn, 2015), but will be discussed more elaborate here.

First, a large sample size might seem problematic. A large sample size leads to small intervals, that is, rather precise estimates. Two situations can arise. Either the estimate is correct or there was bias in the original study. While this may seem problematic, it actually is not. In case the estimate is unbiased, the replicated (difference) in means will be close to the original (difference) in means leading to strong support for the replication hypothesis. However, if bias was present in the original study, the (difference in) means in the replication study will be rather different from the difference specified in the replication hypothesis. This will lead to a rather strong rejection of the replication hypothesis. Both situations give the desired result: either strong support for the replication hypothesis in the absence of bias, or strong support against the replication hypothesis in the presence of bias. Concluding: a large sample size, and bias, are no problem.

Second, a small sample seems problematic. A small sample size leads to relatively big intervals. The bigger an interval, the higher the chances are that the replication hypothesis receives support. One should wonder if it is useful to conduct the replication study when the replication hypothesis is vague and non informative. The result based method should not be used for original studies with too small sample sizes. A critical assessment of the replication hypothesis is needed: do you think it is realistic to reject the replication hypothesis based on the results? If this is not the case, different kinds of hypotheses could be formulated. For instance, the informative hypothesis from the theory based method could be used, even based on the results. That is, the ordering could be specified based on the ordering in the results of the original study. Note that for the theory based method the small sample issue does not arise, since the hypotheses are formulated without involving the sample sizes or reported means and standard deviations from the original study.

To summarize, the theory based method could be used regardless of the properties of the original study. The result based method can be used when the original study has a large sample size, regardless of possible bias. In the case of a small sample in the original study, it is better to avoid the result based method and to use the theory based method. The flexibility in formulating hypotheses ensures that a proper evaluation of the corroboration of the results should be possible.

## 3.4 Conducting the replication study

After determining the hypotheses that will be tested, the replication study can be executed. This includes the set-up of the study, obtaining the participants, and collecting the data. The methodological set-up of the replication study is not within the scope of this paper. For the decisions to be made for the set-up, we refer to articles such as Brandt et al. (2015), Hüffmeier et al. (2016), and Wilson (2016). The three relevant issues to be discussed here are the competing hypothesis, the interpretation of the Bayes factor (BF), and Bayesian updating.

### 3.4.1 The competing hypothesis

The goal of the replication study is to use new data to determine the relative support for the replication hypothesis versus its complement: "not the replication hypothesis". This enables answering the question "Are the results of the original study replicated or not". The complement of an informative hypothesis specifying two group means is easy to understand: the complement of  $\mu_1 < \mu_2$  is simply put  $\mu_1 \geq \mu_2$ . For hypotheses with more groups and interval hypotheses it is harder to capture the complement. Therefore, the notation "not  $H_t$ " is used. For instance, in the theory based method the replication hypothesis specified that the mean of the Distance-group was smaller than the mean of the Intermediate-group which in its turn is smaller than the mean of the Closeness-group ( $H_{t_s}: \mu_{Distance} < \mu_{Intermediate} < \mu_{Closeness}$ ). The competing hypothesis contains every situation where this is not the case (e.g.  $\mu_{Intermediate} < \mu_{Distance} < \mu_{Closeness}$ , but also  $\mu_{Closeness} < \mu_{Intermediate} < \mu_{Distance}$  and more),  $H_c$ : "not  $H_{t_s}$ ". Also for the other hypotheses in the running example  $H_c$  equals "not the replication hypothesis".

### 3.4.2 Comparing the replication hypothesis with its complement

The BF will be used to determine the relative support in the data for the replication hypothesis versus its complement. See for a general introduction of the BF Kass and Raftery (1995), and for a tutorial focused on hypotheses evaluation Hoijtink, Mulder, et al. (2019). Here, the Bayes factor as implemented in the R package `bain` will be used (for the statistical background the interested reader is referred to Gu et al., 2018; Hoijtink, Gu, et al., 2019a; and Hoijtink, Mulder, et al., 2019). We now provide a rather accessible account of this BF. For a more thorough explanation and elaboration of how `bain` calculates the

BF, the interested reader is referred to Appendix 3.A. This BF is calculated by evaluating the fit and the complexity of each hypothesis:

$$BF_{tsc} = \frac{f_t}{c_t} \times \frac{1 - c_t}{1 - f_t}, \quad (3.5)$$

where  $f_t$  denotes the fit of  $H_t$ ,  $c_t$  denotes the complexity of  $H_t$  and  $1 - f_t$  and  $1 - c_t$  denote the fit and complexity of the complement, respectively. When informative and interval hypotheses are evaluated both the fit and the complexity are probabilities. A really badly fitting hypothesis has a fit close to zero, and a very good fitting hypothesis has a fit close to one. A very specific hypothesis has a complexity close to zero, and a hypotheses that is not very specific has a complexity close to one. Note that,  $BF_{rc}$  is obtained if the subscripts  $t$  in Equation 3.5 are replaced by  $r$ .

The BF is a measure of the relative support in the data for  $H_r$  and  $H_c$ . If, for example,  $BF_{rc} = 9$ , there is 9 times more support in the data for  $H_r$  than for  $H_c$ . If  $BF_{rc} = 0.2$  there is 5 times more support for  $H_c$  than for  $H_r$ . As can be seen in Equation 3.5 support is quantified using fit (which should be good) and complexity (which should be small because more specific hypotheses allow for better predictions). The statistical elaboration of fit and complexity can be found in Gu et al. (2018) and Hoijsink, Mulder, et al. (2019). Here only two short and intuitive illustrations will be given. The fit of  $H_{t_s} : \mu_1 > \mu_2 > \mu_3$  is good if  $M_1 = 6, M_2 = 3, M_3 = 1$  because the sample means are in agreement with the order restrictions specified by the hypothesis, and bad if  $M_1 = 5, M_2 = 6, M_3 = 9$ . The complexity of  $H_{t_s} : \mu_1 > \mu_2 > \mu_3$  equals  $1/6$  because one of six possible ordering of three means is specified. The complexity of  $H_{t'_s} : \mu_1 > (\mu_2, \mu_3)$  equals  $2/6$  because two orderings are specified ( $\mu_1 > \mu_2 > \mu_3$  and  $\mu_1 > \mu_3 > \mu_2$ ). Thus,  $H_{t_s}$  is more specific than  $H_{t'_s}$  and therefore it has a smaller complexity.

To compute the complexity a parameter called "fraction" has to be set. This parameter determines the weight of the complexity in the computation of the Bayes factor. The Bayes factor for the evaluation of an informative hypothesis specified using only inequality constraints versus its complement is independent of the choice of this fraction (Hoijsink, Mulder et al., 2019; Mulder, 2014). However, this is not the case for interval hypotheses or informative hypotheses specified using at least one equality constraint. The default setting in **bain** (**fraction** = 1) is chosen such that the Bayes factor tends to prefer the interval hypothesis or an informative hypothesis using at least one equality constraint over its complement unless there is substantial evidence in the data against the

interval or informative hypothesis. There is a good reason to prefer the interval hypothesis or informative hypothesis specified with at least one equality constraint (since these state that there is no relevant effect) over its complement (which states that there is a relevant effect). "In an era of heightened awareness of publication bias, sloppy science, and irreproducibility of research results, researchers should be conservative, that is, convincing evidence is needed before the alternative hypothesis [there is a (relevant) effect] is preferred over  $H_0$  [or there is no relevant effect]" (Hojtink, Mulder, et al., 2019, p. 30). However, when the interval hypothesis or the informative hypothesis represents the results of an original study, the reasoning should be the other way around, that is, convincing evidence is needed before the interval or informative hypothesis is accepted over its complement, that is, the Bayes factor should prefer the complement unless there is substantial evidence in the data in favor of the interval or informative hypothesis. This can be achieved in **bain** using **fraction** = 3. The interested reader is also referred to the section sensitivity analysis in Hoijtink, Mulder, et al. (2019) where this topic is further elaborated.

Note that, the BF does not lead to a dichotomous decision in favor or against the replication hypothesis, it is a measure of support. If  $BF_{rc} = 82$ , it is clear that the replication hypothesis is preferred by the data, if  $BF_{rc} = 9$ , there still is substantial support for the replication hypothesis; if  $BF_{rc} = 3.5$ , there is some support for the replication hypothesis but its complement can not be disqualified; and, if  $BF_{rc} = 1.4$ , the replication data provide about equal support for both hypotheses. For a proper interpretation of the BF, general purpose cut-off values are unnecessary and, in fact, should be avoided (see, for example, the discussion of cut-off values in Hoijtink, Mulder et al., 2019).

Assuming that a priori each hypotheses is equally likely, as is usually done, the BF can be translated into so-called posterior model probabilities (PMP) which quantify the certainty of our preference for one of the hypotheses being tested. For example, if  $BF_{rc} = 4$ ,  $PMP_r = 4/(1 + 4) = .80$  and  $PMP_c = 1/(1+4) = .20$ , which implies that a preference for  $H_r$  comes with a Bayesian error probability of .20. If, for example,  $BF_{rc} = 11$ ,  $PMP_r = 11/(1+11) = .92$  and  $PMP_c = .08$ . This implies that in this situation a preference for  $H_r$  comes with a Bayesian error probability of .08. The interested reader is referred to Hoijtink, Mulder, et al. (2019a) for further elaboration. Note that, in the sequel we will use both BFs and PMPs to quantify the support in the data for the hypotheses entertained.

### 3.4.3 Determining the desired support

How much support we aim for if the replication hypothesis or its complement is true needs to be determined in advance. Sufficient support need not be especially strong (e.g.:  $BF_{rc} = 4$ , or  $BF_{rc} = 6$ ) if the researcher is testing an idea or hypothesis new to their research field and simply wants to know whether it is worthwhile to pursue this theory. However, when established theories are retested and the researcher is thus very certain about the hypothesis of interest, more support would be required (e.g.:  $BF_{rc} = 9$ , or  $BF_{rc} = 11$ ). Each researcher has to decide for the situation at hand how much support is required. In line with the last lines of the previous section, to avoid the introduction of new cut-off values, general purpose guidelines will not be provided. Please note that the desired support is also needed in the next subsection where we will introduce Bayesian updating. Note furthermore, that the use of support should not lead to dichotomous decisions. What it should lead to is a preference for one of the hypotheses involved, where the strength of this preference is quantified using the BF and PMPs. E.g. if  $BF_{rc} = 6$ , there is a preference for  $H_r$ ,  $PMP_r = .86$ , but  $H_c$  can not yet be ignored, since a preference for  $H_r$  comes with a Bayesian error probability of  $(1 - .86) = .14$ . If  $BF_{rc} = 9$ , there is a stronger preference for  $H_r$ , but still  $H_c$  can not completely be ignored: a preference for  $H_r$  comes with a Bayesian error probability of .10.

For the running example, different decisions could be made per hypothesis. For the theory based method ( $H_{ts} : \mu_{Distance} < \mu_{Intermediate} < \mu_{Closeness}$  and both  $H_{t_{min1}}$  and  $H_{t_{min2}}$ ), a support of  $BF_{tc} = 3$  or  $BF_{ct} = 3$  could suffice. From the introduction of Williams & Bargh (2008) it can be seen that they pursue the evaluation of a new theory, therefore one could see this as a first exploration where some support would already be satisfactory. The hypotheses from the results based method ( $H_{rM} : 5.28 < \mu_{Closeness} < 5.94$  &  $4.90 < \mu_{Intermediate} < 5.56$  &  $4.53 < \mu_{Distance} < 5.19$  and  $H_{rD} : -0.09 < \mu_{Closeness} - \mu_{Intermediate} < 0.85$  &  $-0.10 < \mu_{Intermediate} - \mu_{Distance} < 0.84$ ) should receive a substantial amount of support. The data in the original study was very convincing with respect to this hypothesis, therefore it would be reasonable to aim for a BF of 10 in favor of either this replication hypothesis or its complement. Note that, the BF values chosen in this paragraph are, in the role of replication researchers, our choices. Other replication researchers may very well come up with different values.

### 3.4.4 An important property of the replication study

In classical statistics power analyses are often conducted (Cohen, 1988; Erfelder et al., 1996). Power analysis provides the needed sample size to reach meaningful conclusions for hypotheses formulated in standard forms. For Bayesian statistics there is a very good alternative: Bayesian updating (Hojtink, Mulder, et al., 2019; Rouder, 2014). Bayesian updating is conducted by sampling additional data until the minimum amount of support in favor of the replication hypothesis or its complement is obtained or the available resources (time and money) to collect additional data have been exhausted. Of course, replication studies should be adequately powered, see for example Simonsohn (2015) and Morey and Lakens (2016). Therefore, it is recommended to execute some form of Bayesian design analysis in order to determine whether the resources available are sufficient to obtain a "well-powered" replication study. Bayesian design analysis is under development, the interested reader is referred to Schönbrodt and Wagenmakers (2018), and Fu (2022) for the current state of the art.

Practically updating can be executed by starting with a reasonable sample size, e.g.  $n = 20$  per group. After the first round of data collection the BF is determined. If this BF meets the desired support, see the previous subsection, data collection is finished and the research process moves on. If the BF does not meet the desired support for either hypothesis, more data is collected. After each round of data collection (with for instance  $n = 20$  per group) the BF is determined. Once it reaches the desired amount of support for either hypothesis the data collection stops. Please note that when the desired BF is relatively small, e.g.  $BF_{rc} = 4$  (which corresponds to PMPs of .80 and .20 for  $H_r$  and  $H_c$ , respectively), the probability that the best hypothesis is incorrectly preferred is still relatively large. On the other hand, if the desired BF is relatively large, e.g.  $BF_{rc} = 11$ , the probability of an incorrect preference is relatively small, that is, if  $H_r$  is preferred the error probability, that is, the PMP associated with  $H_c$ , equals .08. Updating cannot be illustrated here, since this paper is based on existing original and replication studies and therefore the sample sizes are fixed. However, we will provide one example where updating is applied using additional hypothetical data and mention for the other hypotheses what decision would be made with regards to updating.

### 3.4.5 Evaluating the replication hypothesis using the replication data

To calculate  $BF_{tc}$  or  $BF_{rc}$  the R package `bain` (Hoijsink, Mulder, et al., 2019; <https://informative-hypotheses.sites.uu.nl/software/bain/>) is used. The codes and data used for the analyses presented in this paper can be found on the OSF page for this study: <https://osf.io/up3rv/>.

#### Theory based method, $H_{ts}$

For  $H_{ts} : \mu_{Distance} < \mu_{Intermediate} < \mu_{Closeness}$ , the desired support was  $BF = 3$  for either hypothesis. In the replication sample there were 44 and 38 participants in the groups. The data eventually provided minimal support for the replication hypothesis,  $BF_{tsc} = 2.10$ , but support does not reach the desired level. The difference between the means of the Distance group ( $M = 5.31$ ,  $SD = 1.15$ ), the Intermediate group ( $M = 5.31$ ,  $SD = 1.07$ ), and the mean of the Closeness group ( $M = 5.44$ ,  $SD = 0.83$ ) is not large enough (Cohen's  $d = 0.13$  between both Distance and Closeness, and between Intermediate and Closeness), nor is the ordering convincingly out of order, to render the required evidence in favor of either the replication hypothesis or its complement.

Therefore, Bayesian updating (Rouder, 2014) can be applied here to obtain stronger support for either the replication hypothesis or its complement. Since we are not able to collect more data ourselves, we duplicate the dataset of the replication study. *This is not a recommended procedure!* It is used here only to demonstrate how updating works. We will assume that we have collected additional data with coincidentally the same sample means, standard deviations, and sample size as the first half. The  $BF_{tsc}$  is calculated for the updated dataset. Now, the result is  $BF_{tsc} = 2.66$  (versus the previous  $BF_{tsc} = 2.10$ ). The extended dataset gives almost 3 times more support for the replication hypothesis than for the complement. Since the desired amount of support is still not achieved, it is reasonable to continue collecting data until the threshold of  $BF = 3$  is achieved in favor of the replication hypothesis or its complement. *Note again that in a real situation the researcher would need to actually collect more data for Bayesian updating, which we did not do.*

After adding more data, in this case triplicating the original data,  $BF_{tsc}$  is calculated again. Now, the result is  $BF_{tsc} = 3.10$  (versus the previous  $BF_{tsc} = 2.66$ ). The extended dataset gives more than three times as much support for the replication hypothesis over the complement. Now, the desired amount of support is achieved, we have a preference for  $H_{ts}$ , however this preference comes

with an error probability equal to  $\text{PMP}_c = .24$ .

### Theory based method, $H_{t_{min1}}$ and $t_{min2}$

For the minimal difference hypothesis two hypotheses were formulated:  $H_{t_{min1}}$  ( $\mu_{Distance} + 0.41 < \mu_{Intermediate} + 0.19 < \mu_{Closeness}$ ) and  $H_{t_{min2}}$  ( $\mu_{Distance} + 0.21 < \mu_{Intermediate} + 0.10 < \mu_{Closeness}$ ). Both hypotheses were formulated to show the impact of interpreting the minimal difference of 0.2 SD: either between adjacent groups or between the most extreme group means. A desired support of  $\text{BF} = 3$  was set for this hypothesis. With approximately  $n = 40$  per group we can see that the actual difference is not in agreement with the hypothesis. The mean of the Closeness group is 0.13 higher than both the mean of the Distance and the mean of the Intermediate group. Only the difference between the means of the Intermediate and the Distance group are according to the situation describe in  $H_{t_{min1}}$ . All other means are not in accordance with either hypothesis. The resulting BF's were:  $\text{BF}_{t_{min1c}} = 0.57$  and  $\text{BF}_{t_{min2c}} = 0.11$ . There is no convincing support to prefer  $H_{t_{min1}}$  or its complement:  $\text{PMP}_{t_{min1}} = .36$  and  $\text{PMP}_c = 0.64$ , however, updating could be used to reach the desired level of support. There is clear support for  $H_c$  over  $H_{t_{min}}$ ,  $\text{BF}_{t_{minc}} = .11$ , that is,  $\text{BF}_{ct_{min}} = 9.10$  with  $\text{PMP}_{t_{min}} = .10$  and  $\text{PMP}_c = .90$ .

### Results based method, $H_{r_M}$

For the second method we consider the hypothesis  $H_{r_M}$  ( $H_{r_M} : 5.28 < \mu_{Closeness} < 5.94$  &  $4.90 < \mu_{Intermediate} < 5.56$  &  $4.53 < \mu_{Distance} < 5.19$ ). A desired support of  $\text{BF} = 10$  was set for this hypothesis. The intervals appear small enough to formulate meaningful conclusions. Some overlap between adjacent means is present, but the extremes do not overlap. Furthermore, the interval cover around 10% of the scale, which leaves enough room for scores outside the intervals. We can continue with the results based method. With approximately  $n = 40$  per group, two out of three means ( $M_{Closeness} = 5.44, SD = 0.83$  &  $M_{Intermediate} = 5.31, SD = 1.15$ ) from the replication study fall within their respective confidence intervals from the hypothesis. The mean of the Distance group ( $M = 5.31, SD = 1.07$ ) does not fall within its hypothesized confidence interval. This results in some support for the replication hypothesis,  $\text{BF}_{r_Mc} = 1.92$ . There is some reason to further explore  $H_{r_M}$ , though there is no convincing support according to the standards set by us to prefer the presented hypothesis or its complement, the Bayesian error probability associated with a preference for  $H_{r_M}$  is  $\text{PMP}_c = .34$ . Updating could, in principle, be used to obtain stronger

evidence for either hypothesis.

### Results based method, $H_{r_D}$

Lastly the interval hypothesis specifying the difference between the means ( $H_{r_D} : -0.09 < \mu_{Closeness} - \mu_{Intermediate} < 0.85 \ \& \ -0.10 < \mu_{Intermediate} - \mu_{Distance} < 0.84$ ) is considered. For this hypothesis a desired support of  $BF = 10$  was set. The difference scores are measured on a scale from -6 to 6. This means that the intervals cover about 10% of the scale, which leaves enough room for scores outside the intervals. We can continue with the results based method. Both intervals include the actual difference between the groups means found in the replication study, this results in  $BF_{r_{DC}} = 4.44$ . There is support although not convincing enough for our goals to prefer either the replication hypothesis nor its complement. To achieve this level of support, updating could be used.

### 3.4.6 Conclusion

In reality, researchers would evaluate one replication hypothesis. That we evaluated four hypotheses was only to show the versatility of our approach. But it may be clear that most BFs show no clear preference for either hypothesis. The only BF that gives a clear answer shows strong support for the complement. All in all, we conclude that the results of Williams and Bargh (2008) are not convincingly corroborated by Joy-Gaba et al. (2012).

## 3.5 Example 2 - Janssen, Schirm, Mahon, and Caramazza

For the second example, the original article of Janssen et al. (2008) and the replication of Study 1 in this paper by Galak (2012) will be used. The decision was made to conduct the replication using the simple ordering hypothesis from the theory based method, because the goal of this replication is to test the general conflicting theories presented in the introduction of Janssen et al. (2008). This second example will illustrate a situation where competing theories are tested against each other, as well as a hypothesis specified using an equality constraint and an inequality constraint.

### **3.5.1 Step 1 - Reading the introduction section**

#### **Reading the introduction section**

The goal of the study is to test whether the semantic interference effect arises in a delayed naming task. Participants were shown 20 pictures of common objects, half with low frequency names (used 1 to 9 times per million words) and half with high frequency names (used 72 to 724 times per million). In each trial, one of the 20 pictures (e.g., car) was shown on the screen. After 1000 ms, a distractor word was shown beneath the picture. The distractor could be either related (e.g., truck) or unrelated (e.g., table). Participants were instructed to name the picture as soon as the distractor word appeared on the screen. The semantic interference effect claims that a related distractor word makes it easier to find the correct word, so the response time is shorter when the distractor word is related. Since the participant sees the picture for some time, it is expected that the frequency (1-9 times or 72-724 times per million) has no effect on the time to give the proper response. If the semantic interference effect arises at a post-lexical level of processing, that is, after thinking about what to say, the distractor word would influence reaction time. The researchers wanted to investigate whether this effect appears in such a delayed naming task.

### **3.5.2 Step 2 - Coding of statements in the introduction section**

From the introduction section a number of statements were coded that related to the hypotheses of interest to the researchers of the original paper:

- "The primary source of empirical evidence cited in support of lexical selection by competition is the semantic interference effect (...): Naming a picture of an object (e.g., CAR) is slower in the context of a semantic category coordinate distractor word (e.g., truck) compared to an unrelated distractor word (e.g., table)." (Janssen et al., 2008, p.2)
- "If the semantic interference effect arises at the level of lexical selection, then the semantic interference effect will not be observed when participants delay their picture naming responses." (Janssen et al., 2008, p.3)
- "However, if the semantic interference effect arises at a postlexical level of processing, then semantic interference should be observed in a delayed naming task." (Janssen et al., 2008, p.3)

- "If participants have already retrieved the lexical representations corresponding to the target picture names in the delayed naming condition at the time the cue is presented, then there should be no effect of the frequency of the target pictures on naming latencies." (Janssen et al., 2008, p.4)

### 3.5.3 Step 3 - Reading the methods & results section

The methods section makes clear that a 2 (Frequency: high; HF, vs low; LF) x 2 (Relatedness: related; rel, vs unrelated; unrel) ANOVA design was used. The dependent variable was the time the participants needed to name the object.

### 3.5.4 Step 4 - Coding of statements in the methods & results section

From the results section, the three statements with respect to hypothesis testing were coded. Note that the interest is on what is being tested, not on the results of the test:

- "The R[esponse] T[ime] analysis revealed a main effect of semantic relatedness, (...) ;  $F_2(1, 38) = 4.4, MSE = 1,137.2, p < .05$ , indicating slower RTs in the semantically related than the unrelated condition." (Janssen et al., 2008, p.6)
- Neither the main effect of frequency nor (...) reached significance (all  $F_s < 1$ ). (Janssen et al., 2008, p.6)
- Neither (...) nor the interaction between frequency and semantic relatedness reached significance (all  $F_s < 1$ ). (Janssen et al., 2008, p.6)

### 3.5.5 Step 5 - Selecting the statements from the introduction section that are actually tested

Based on the statements from the results section, it appeared that the first statement was not tested, it was merely a description of the effect under investigation in statements two and three. However, Statements 2 and 3 were tested.

### 3.5.6 Step 6 - Translating statements to informative hypotheses

Statements remaining from Step 5 will now be translated into simple ordering hypotheses:

- "If the semantic interference effect arises at the level of lexical selection, then the semantic interference effect will not be observed when participants delay their picture naming responses." can be represented as:

$$\mu_{1A} + \mu_{2A} = \mu_{1B} + \mu_{2B}$$

- "However, if the semantic interference effect arises at a postlexical level of processing, then semantic interference should be observed in a delayed naming task." can be represented as:

$$\mu_{1A} + \mu_{2A} > \mu_{1B} + \mu_{2B}$$

- "If participants have already retrieved the lexical representations corresponding to the target picture names in the delayed naming condition at the time the cue is presented, then there should be no effect of the frequency of the target pictures on naming latencies." can be represented as:

$$\mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$$

The first two statements describe the same main effect, but with different outcomes. One suggests that there is a main effect: the reaction time is lower with unrelated distractors than with related distractors, the other describes the absence of this effect: both groups score equally on average. The third statement describes the other main effect, that of frequency. This results in the following replication hypotheses:

$$H_{t_1} : \mu_{1A} + \mu_{2A} = \mu_{1B} + \mu_{2B} \ \& \ \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B},$$

$$H_{t_2} : \mu_{1A} + \mu_{2A} > \mu_{1B} + \mu_{2B} \ \& \ \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B},$$

using the replication data, each can be tested against its complement, and, additionally, both can be tested against each other.

### 3.5.7 Determining the desired support

From the literature it is apparent that there is a clear and established theory regarding the effect of frequency on the reaction time. For the relatedness of

distractors, there are contradicting ideas and theories. For this reason, the desired support is set to 5 for both of the replication hypotheses versus their respective complements. However, the replication hypotheses can also be tested against each other, and to gather some idea on which hypothesis needs further exploration, here a minimum BF of 5 is also chosen.

### 3.5.8 Calculate *BF* for the replication data

Three BFs are calculated: each replication hypotheses versus its complement and the replication hypotheses versus each other. Note that, for the same reason highlighted for interval hypotheses, here too `bain` was used with `fraction = 3` for all analyses. The statistics for both the original and the replication study are displayed in Table 3.3.

- $H_{t_1} : \mu_{1A} + \mu_{2A} = \mu_{1B} + \mu_{2B} \ \& \ \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$  versus its complement. The result of the analysis,  $BF_{t_1c} = 4.57$ , indicates a slight preference for the first replication hypothesis, the error probability associated with a preference for  $H_{t_1}$  equals  $PMP_c = .18$ . This is corroborated by the  $\eta^2$ 's for both main effects that can be found in Table 3.3, it is .02 for the first part of  $H_{t_1}$  and .03 for the second part, that is, both are rather small.
- $H_{t_2} : \mu_{1A} + \mu_{2A} > \mu_{1B} + \mu_{2B} \ \& \ \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$  versus its complement. The result of this analysis,  $BF_{t_2c} = 2.24$ , lends hardly any support for the replication hypothesis versus its complement, the error probability associated with a preference for  $H_{t_2}$  equals  $PMP_c = .31$ .
- $H_{t_1}$  versus  $H_{t_2}$ . When the two replication hypotheses are tested versus each other, the support for the first is hardly larger than for the second,  $BF_{t_1t_2} = 2.04$ , the error probability associated with a preference for  $H_{t_1}$  equals  $PMP_{t_2} = .33$ .

**Table 3.3:** Descriptive statistics and results of the original study of Janssen, Schirm, Mahon, & Caramazza (2008) and its replication by Galak (2012).

|                                                                                                     | Original study by Janssen,<br>Schrim, Mahon, & Caramazza (2008) |       |       |       | Replication study by Galak (2012) |                  |                  |                  |
|-----------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|-------|-------|-------|-----------------------------------|------------------|------------------|------------------|
|                                                                                                     | Group                                                           | Group | Group | Group | Group                             | Group            | Group            | Group            |
|                                                                                                     | 1A                                                              | 1B    | 2A    | 2B    | 1A                                | 1B               | 2A               | 2B               |
| $M (SD)$                                                                                            | nr <sup>1</sup>                                                 | nr    | nr    | nr    | 824.31 (270.71)                   | 815.14 (255.35)  | 834.00 (273.51)  | 831.27 (267.87)  |
| 95% CI around $M$                                                                                   |                                                                 |       |       |       | (705.67, 942.95)                  | (703.23, 927.05) | (714.13, 953.87) | (713.87, 948.67) |
| $n$                                                                                                 | nr                                                              | nr    | nr    | nr    | 20                                | 20               | 20               | 20               |
| $H_0 : \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$                                                   | $p > .05$                                                       |       |       |       | $p = .27, \eta^2 = .03$           |                  |                  |                  |
| $H_0 : \mu_{1A} + \mu_{2A} = \mu_{1B} + \mu_{2B}$                                                   | $p < .05, \eta^2 = nr$                                          |       |       |       | $p = .43, \eta^2 = .02$           |                  |                  |                  |
| $H_0 : \mu_{1A} - \mu_{1B} = \mu_{2A} - \mu_{2B}$                                                   | $p > .05$                                                       |       |       |       | $p = .64, \eta^2 = .002$          |                  |                  |                  |
| $H_{t_1} : \mu_{1A} + \mu_{2A} = \mu_{1B} + \mu_{2B}$ & $\mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$ |                                                                 |       |       |       | $BF_{t_1c} = 4.57$                |                  |                  |                  |
| $H_{t_2} : \mu_{1A} + \mu_{2A} > \mu_{1B} + \mu_{2B}$ & $\mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$ |                                                                 |       |       |       | $BF_{t_2c} = 2.24$                |                  |                  |                  |
|                                                                                                     |                                                                 |       |       |       | $BF_{t_1t_2} = 2.04$              |                  |                  |                  |

Note. <sup>1</sup> = This information is not reported (nr), nor can it be calculated based on the reported information. 2. The means are all measured in milliseconds.

### 3.5.9 Interpret outcome

We chose 5 as the desired support for both replication hypotheses versus their respective complements and versus each other.  $H_{t_1}$  ( $\mu_{1A} + \mu_{2A} = \mu_{1B} + \mu_{2B}$  &  $\mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$ ) received more than 4 times as much support as its complement.  $H_{t_1}$  receives about double the support as  $H_{t_2}$  ( $\mu_{1A} + \mu_{2A} > \mu_{1B} + \mu_{2B}$  &  $\mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$ ). Both of these BFs do not meet the desired support. There is some indication that the original results of Janssen et al. (2008) formulated in terms of  $H_{t_1}$  were corroborated by Galak (2012). However, since the sample sizes of the replication study were relatively small (20 persons per group), and the desired support for either hypothesis was not achieved, it is relevant to consider collecting additional data for the replication study and, subsequently, to update the Bayes factors.

## 3.6 Discussion

Inspired by the reproducibility project psychology (OSC, 2012; 2015) two refined methods were offered to plan and evaluate replication studies. In the theory based method the replication hypothesis was formulated based on the introduction of the original paper. This method would be executed when a theory elaborated in the introduction of the original study is evaluated using a replication study. Two kinds of informative hypotheses can be formulated here, a simple ordering hypothesis or a minimal difference hypothesis. In the results based method, the results of the original study are used to formulate the replication hypothesis. Here, an interval hypothesis around the means or around the difference between means is formulated.

We believe both methods enable replication studies that meet the needs and desires of replicating researchers. They can determine how much the results of a replication study favor or contradict the original, using the BF. Where necessary, Bayesian updating allows for the collection of data until the desired degree of certainty is obtained. Our methods allow to formulate multiple concurrent hypotheses as was illustrated in Example 2, and provide flexibility in formulating the replication hypothesis. Our approach is a remedy against the "null-ritual" Gigerenzer (2004) and explicitly addresses the importance of testing the correct hypothesis (which is not always done, Cho & Abe, 2013).

For the evaluation of both methods some subjective decisions had to be made. For both methods the desired support has been set to  $BF = 3$  or 10. For the theory based method one has to decide whether the minimal difference

applies to each pair of group or to the extreme groups. For the results based method `fraction = 3` has been chosen. For the results based method we used 95%-intervals, 90% or 99% intervals could also have been used. For the results based method one has to decide whether to focus on means or a difference between means. Within this article no investigation on the impact of the choices has been executed. The focus in this paper is on the hypotheses one formulates. It is essential to test the correct hypothesis, and the main message from this study is how to formulate that hypothesis. Therein still lies a subjective decision, replicating researchers should critically assess when a result corroborates the original study. In some instances the difference between adjacent means is important, sometimes a focus on the difference between the smallest and the largest means is most relevant (for the theory based method). And whereas in some circumstances hypotheses should be based on the means, in other one prefers the difference between means. Researchers could critically assess which situation applies to their situation before conducting the replication study. A sensitivity analysis could be helpful to understand the effect of these types of decisions. The remainder of the paper explains the further steps necessary to analyse the replication hypothesis. During that process, for each of the subjective decisions the mainstream options were followed.

A major strength of these methods is the possibility of updating (Rouder, 2014). Updating makes sure that studies end with convincing results. No estimates, assumptions, or guesses are necessary, unlike for power analyses. The replicating researchers only need to determine what the stopping point is, i.e. what the desired amount of support is.

The use of the BF brings one more feature to the table which is not yet discussed: information synthesis (Kuiper et al., 2013). Though not exclusive to the BF, synthesis is very easily conducted with the BF. With information synthesis, multiple BFs can be combined to create a meta view on the support for a certain hypothesis. It is necessary that the BFs under consideration are calculated under the same conditions, that is, with the same hypotheses. In the case of the theory based method, this leads to an interesting possibility. If the introduction section of the original paper is constructed before collecting the data for that paper, there are two datasets (the original and the replication) able to provide information on the replication hypothesis. By calculating the  $BF_{tc}$  twice, and multiplying the two  $BF_{tc}$ s a new overall  $BF_{tc}$  is generated, combining the information from both the original and the replicating study. For the results based method this is not possible, since the replication hypothesis is based on the results of the original paper. This leads to an interesting discussion for

theory based method: can we assume that the hypotheses from the introduction section are not influenced by the data from that very same paper? Ideally; this assumption should be met, but is that also true in reality?

The approaches presented in this paper can easily be executed using the R package `bain` (<https://informative-hypotheses.sites.uu.nl/software/-bain/>). The approach presented in this paper is therefore well within reach of psychological researchers that want to execute a replication study in order to evaluate the results of an original study.

## Bibliography

- Anderson, S., Kelley, K., & Maxwell, S. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, 28, 1547–1562. <https://doi.org/10.1177/0956797617723724>
- Anderson, S., & Maxwell, S. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21, 1–12. <https://doi.org/10.1037/met0000051>
- Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of personality and social psychology*, 71(2), 230–244. <https://doi.org/10.1037/0022-3514.71.2.230>
- Boeije, H. (2010). *Analysis in qualitative research*. SAGE publications Ltd. <https://doi.org/10.5785/26-2-24>
- Brandt, M., IJzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R., Grange, J., Perugini, N., Spies, J., & Van 't Veer, A. (2015). 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>
- Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.
- Cho, H.-C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, 66. <https://doi.org/10.1016/j.jbusres.2012.02.023>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.
- Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5, 1664–1078. <https://doi.org/10.3389/fpsyg.2014.00781>

- Doyen, S., Klein, O., Pichon, C., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS one*, 7(1). <https://doi.org/10.1371/journal.pone.0029081>
- Erdfelder, E., Faul, F., & Buchner, A. (1996). Gpower: A general power analysis program. *Behavior Research Methods, Instruments, Computers*, 28, 1–11. <https://doi.org/10.3758/BF03203630>
- Etz, A., & Vandekerckhove, J. (2016). A bayesian perspective on the reproducibility project: Psychology. *PLoS one*, 11(2). <https://doi.org/10.1371/journal.pone.0149794>
- Field, S., Hoekstra, R., Brinkmann, L., & Van Ravenzwaaij, D. (2019). When and why to replicate: As easy as 1, 2, 3? *Collabra: Psychology*, 5(46). <https://doi.org/10.1525/collabra.218>
- Fu, Q. (2022). *Sample size determination for bayesian informative hypothesis testing*. [doctoral dissertation] [Utrecht University Repository. <https://dspace.library.uu.nl/handle/1874/416118>]. Utrecht University.
- Galak, J. (2012). Replication of study 1 by janssen, schirm, mahon, caramazza (2008, jep:lmc) [Retrieved from <https://osf.io/uhpyr/>].
- Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, 60(4), 328–331. <https://doi.org/10.1198/000313006X152649>
- Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. In *Blackwell handbook of judgment and decision making*. Blackwell Publishing. <https://doi.org/10.1002/9780470752937>.
- Glaser, B. (1978). *Theoretical sensitivity*. Mill Valley, CA: The Sociology Press.
- Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximate adjusted fractional bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, 71, 229–261. <https://doi.org/10.1111/bmsp.12110>
- Harms, C. (2018). A bayes factor for replications of anova results. *The American Statistician*, 73, 327–339. <https://doi.org/10.1080/00031305.2018.1518787>
- Hawthorne, J. (2021). Inductive logic [Retrieved December 2, 2022. The Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/archives/spr2021/entries/logic-inductive/>].
- Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. New York: Chapman; Hall/CRC.
- Hoijtink, H., Gu, X., Mulder, J., & Rosseel, Y. (2019a). Bayesian evaluation of informative hypotheses for multiple populations. *British Journal of*

- Mathematical and Statistical Psychology*, 72, 219–243. <https://doi.org/10.1111/bmsp.12145>
- Hojtink, H., Gu, X., Mulder, J., & Rosseel, Y. (2019b). Computing bayes factors from data with missing values. *Psychological Methods*, 24, 253–268. <https://doi.org/10.1037/met0000187>
- Hojtink, H., Mulder, J., Van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the bayes factor. *Psychological Methods*, 24, 539–556. <https://doi.org/10.1037/met0000201>
- Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Psychology*, 66. <https://doi.org/10.1016/j.jesp.2015.09.009>
- Janssen, N., Schirm, W., Mahon, B., & Caramazza, A. (2008). Semantic interference in a delayed naming task: Evidence for the response exclusion hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 249–256. <https://doi.org/10.1037/0278-7393.34.1.249>
- Joy-Gaba, J., Clay, R., & Cleary, H. (2012). Replication of keeping one's distance: The influence of spatial distance cues on affect and evaluation by williams & bargh (2008, psychological science) [Retrieved from <https://osf.io/vnsgq/>].
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795. <https://doi.org/10.2307/2291091>
- Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, 10, 477–493. <https://doi.org/10.1037/1082-989X.10.4.477>
- Kuiper, R., Hoijtink, H., Buskens, V., & Raub, W. (2013). Combining statistical evidence from several studies: A method using bayesian updating and an example from research on trust problems in social and economic exchange. *Sociological Methods and Research*, 42, 60–81. <https://doi.org/10.1177/0049124112464867>
- Lakens, D., Scheel, A., & Isager, P. (2013). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1, 259–269. <https://doi.org/10.1177/2515245918770963>
- Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2018). Replication bayes factors from evidence updating. *Behavior Research Methods*, 51, 2498–2508. <https://doi.org/10.3758/s13428-018-1092-x>
- Marsman, M., Schönbrodt, F., Morey, R., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A bayesian bird's eye view of 'replications of important

- results in social psychology'. *Royal Society Open Science*, 4, 160426. <https://doi.org/10.1098/rsos.160426>
- Morey, R., & Lakens, D. (2016). Why most of psychology is statistically unfalsifiable.
- Morey, R., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72, 6–18. <https://doi.org/10.1016/j.jmp.2015.11.001>
- Mulder, J. (2014). Prior adjusted default bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics and Data Analysis*, 71, 448–463. <https://doi.org/10.1016/j.csda.2013.07.017>
- Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6), 657–660. <https://doi.org/10.1177/1745691612462588>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, 6251. <https://doi.org/10.1126/science.aac4716>
- Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? a statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11(4), 539–544. <https://doi.org/10.1177/1745691616646366>
- Popper, K. (1963). Science as falsification. *Conjectures and refutations*, 1, 33–39.
- Rosenthal, R. (1979). The "file drawer problem" and tolerance for null effects. *Psychological Bulletin*, 86, 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Rouder, J. (2014). Optional stopping: No problem for bayesians. *Psychonomic Bulletin Review*, 21, 301–308. <https://doi.org/10.3758/s13423-014-0595-4>
- Schönbrodt, F., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin Review*, 25, 128–142. <https://doi.org/10.3758/s13423-017-1230-y>
- Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5), 559–569. <https://doi.org/10.1177/0956797614567341>
- Van Aert, R., & Van Assen, M. (2017). Bayesian evaluation of effect size after replicating an original study. *PLoS one*, 12. <https://doi.org/10.1371/journal.pone.0175302>

- Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. <https://doi.org/10.1037/a0036731>
- Williams, L., & Bargh, J. (2008). Keeping one's distance: The influence of spatial distance cues on affect and evaluation. *Psychological Science*, *19*, 302–308. <https://doi.org/10.1111/j.1467-9280.2008.02084.x>
- Wilson, A. (2016). Exact replications in an inexact context: Commentary on ebersole et al. *Journal of Experimental Social Psychology*, *67*, 84–85. <https://doi.org/10.1016/j.jesp.2015.12.008>
- Zondervan-Zwijnenburg, M., van der Schoot, R., & Hoijsink, H. (2019). Testing anova replications by means of the prior predictive p-value. [2019, January 28]. <https://doi.org/10.31234/osf.io/6myqh>

## Appendix 3.A: The Approximate Adjusted Fractional Bayes Factor

### Appendix 3.A: The Approximate Adjusted Fractional Bayes Factor

The Approximate Adjusted Fractional Bayes Factor (AAFBBF) is implemented in the R package `bain` and was developed such that it can be applied in a wide range of statistical models for the evaluation of informative hypotheses in addition to the traditional null and alternative hypotheses. The statistical background of the AAFBBF can be found in (Gu et al., 2018; Hoijtink, Gu, & Mulder, 2019a; Hoijtink, 2022; Hoijtink, Gu, Mulder, & Rosseel, 2019b), more practically oriented tutorials about the use of the AAFBBF can be found in (Hoijtink, Gu, & Mulder, 2019a; Van Lissa et al., 2020). In this appendix the AAFBBF, simplified to the form it attains in the context of ANOVA models, will be presented.

The Bayes factor of  $H_i$ , which denotes either a null ( $H_0$ ) or an informative hypothesis ( $H_i$ ) versus the unconstrained hypothesis  $H_u$  can be written as the ratio of two marginal likelihoods:

$$BF_{iu} = \frac{m(D|H_i)}{m(D|H_u)} = \frac{f_i}{c_i}, \quad (3.6)$$

where  $D$  denotes the data, that is, for an ANOVA, group membership and score on the dependent variable for each person. As was shown by Klugkist, Laudy, and Hoijtink (2005), this ratio of two marginal likelihoods can be rewritten in term of the ratio of the fit  $f_i$  and complexity  $c_i$  of  $H_i$ .

The "approximate" in AAFBBF results from the fact that the fit is computed with respect to a normal approximation of the posterior distribution of  $\mu = [\mu_1, \dots, \mu_G]$ , where  $G$  denotes the number of groups in the ANOVA:

$$g(\mu|D) = \mathcal{N}(\mu|\hat{\mu}, \Sigma), \quad (3.7)$$

where  $\hat{\mu}$  denotes estimates of the means, and  $\Sigma$  the covariance matrix of the esti-

mates which is diagonal with elements  $\sigma^2/N_g$  for  $g = 1, \dots, G$ , where  $\sigma^2$  denotes the pooled within groups variance, and  $N_g$  the sample size in group  $G$ . Already for relatively small sample size per group (e.g., 20) the normal approximation becomes virtually indistinguishable from the true (a t-distribution) posterior distribution of the means. The fit is

$$f_i = \int_{\mu} H_i g(\mu|D) d\mu. \quad (3.8)$$

It is illustrative to note that for hypotheses constructed using only inequality constraints, the fit is the proportion of the posterior distribution in agreement with  $H_i$ . In general it holds that the larger the fit the larger the support in the data for  $H_i$ .

The complexity is based on the "adjusted" "fractional" prior distribution corresponding the posterior distribution:

$$h(\mu|D) = \mathcal{N}(\mu|\mu_{adj}, \Sigma_{frac}), \quad (3.9)$$

where the adjusted prior mean  $\mu_{adj}$  is a value on the boundary of  $H_i$  and its complement  $H_c$ . For example, the complement of  $H_0 : \mu_1 - \mu_2 = 0$  is  $H_c : \text{"not } H_i\text{"}$  and therefore any  $\mu_{1,adj} = \mu_{2,adj}$  can be used as the prior mean. Analogously, a value on the boundary of  $\mu_1 > \mu_2 > \mu_3$  and its complement would be  $\mu_{1,adj} = \mu_{2,adj} = \mu_{13,adj}$ . For interval hypotheses like  $2 < \mu < 4$ ,  $\mu_{adj}$  is set equal to the mean of the interval, that is, for setting the mean of the prior distribution this hypothesis is treated analogously as  $\mu = 3$ . The prior covariance matrix is based on a fraction of the information in the data with respect to the  $\mu$ 's, that is,  $\Sigma_{frac} = \Sigma * 1/b$ , where  $b = [b_1, \dots, b_G]$  with  $b_g = 1/G * J/N_g$ , denotes the fraction of information in group  $G$  used to specify the prior variance of the mean in group  $g$ . The parameter  $J$  will be elaborated in the next paragraph. The complexity is

$$c_i = \int_{\mu \in H_i} h(\mu|D) d\mu. \quad (3.10)$$

It is illustrative to note that for hypotheses constructed using only inequality constraints, the complexity is the proportion of the prior distribution in agreement with  $H_i$ . This implies that the Bayes factor of an informative hypothesis

versus its complement can be written as:

$$BF_{ic} = \frac{f_i}{c_i} / \frac{1 - f_i}{1 - c_i}. \quad (3.11)$$

In general it holds that the larger the complexity the less specific, that is, the less parsimonious  $H_i$ .

The prior means are chosen in the manner elaborated in the previous paragraph to ensure that both  $BF_{0u}$  and  $BF_{ic}$  are consistent (Hojtink, Gu, & Mulder, 2019a; Gu et al., 2018), that is, if the sample size increases these Bayes factors go to 0 if  $H_u$  or  $H_c$  is the best hypothesis and to infinity if  $H_0$  or  $H_i$  is the best hypothesis. This holds for any value of  $J$ . The prior variances are chosen in agreement with the minimal training sample principle (see, for example, Berger & Pericchi, 2004; O'Hagan, 1995), that is, what is the smallest sample size that allows estimation of the parameters of the ANOVA model (the answer is  $G + 1$ ). In the AAFBF  $J$  is chosen not equal to the minimal training sample size, but equal to the number of independent constraints used to specify the hypothesis of interest. In case of  $H_0 : \mu_1 = \mu_2 = \mu_3$  the number of independent constraints equals two, the same holds for  $H_i : \mu_1 > \mu_2 > \mu_3$ . Finally note, that the choice of  $J$  is only relevant if the goal is to evaluate a hypothesis specified using equality or inequality constraints, in all other cases (hypotheses specified using only inequality constraints that are not interval hypotheses) the Bayes factor is completely insensitive to the specification of  $J$  (Mulder, 2014 ; Hoijtink, Mulder, et al., 2019).

## Bibliography

- Berger, J., & Pericchi, L. (2004). Training samples in objective bayesian model selection. *The Annals of Statistics*, *32*, 841–869. <https://doi.org/10.1214/009053604000000229>
- Gu, X., Mulder, J., & Hoijsink, H. (2018). Approximate adjusted fractional bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*, 229–261. <https://doi.org/10.1111/bmsp.12110>
- Hoijsink, H. (2022). Prior sensitivity of null hypothesis bayesian testing. *Psychological Methods*, *27*(5), 804–821. <https://doi.org/10.1037/met0000292>
- Hoijsink, H., Gu, X., Mulder, J., & Rosseel, Y. (2019a). Bayesian evaluation of informative hypotheses for multiple populations. *British Journal of Mathematical and Statistical Psychology*, *72*, 219–243. <https://doi.org/10.1111/bmsp.12145>
- Hoijsink, H., Gu, X., Mulder, J., & Rosseel, Y. (2019b). Computing bayes factors from data with missing values. *Psychological Methods*, *24*, 253–268. <https://doi.org/10.1037/met0000187>
- Hoijsink, H., Mulder, J., Van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the bayes factor. *Psychological Methods*, *24*, 539–556. <https://doi.org/10.1037/met0000201>
- Klugkist, I., Laudy, O., & Hoijsink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, *10*, 477–493. <https://doi.org/10.1037/1082-989X.10.4.477>
- Mulder, J. (2014). Prior adjusted default bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics and Data Analysis*, *71*, 448–463. <https://doi.org/10.1016/j.csda.2013.07.017>
- O’Hagan, A. (1995). Fractional bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 99–138.

- Van Lissa, C., Gu, X., Mulder, J., Rosseel, Y., Van Zundert, C., & Hoijsink, H. (2020). Teacher's corner: Evaluating informative hypotheses using the bayes factor in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(2), 292–301. <https://doi.org/10.1080/10705511.2020.1745644>



## CHAPTER 4

### Applying qualitative methods to experimental designs: A tutorial for the behavioral sciences

Leplaa, H.J., Tönjes, J.A., Bouterse, M., Soppe, K.F.B. & Klugkist, I. (2025). Applying qualitative methods to experimental designs: A tutorial for the behavioral sciences. *PLOS ONE*, 20(6): e0324936. <https://doi.org/10.1371/journal.pone.0324936>

#### Abstract

Studies with experimental designs are almost invariably evaluated with quantitative outcomes and methods, both in behavioral sciences and other disciplines. We argue that there can be added value of using qualitative methods for the evaluation of (behavioral) experiments. Incorporating qualitative data can enhance the ecological validity of a study, by acquiring a more holistic understanding of the phenomenon of interest. There is, however, little methodological guidance on how to implement such an approach.

In this paper we present the different steps and considerations for a qualitative evaluation of results in experimental designs. Methodological guidelines are offered for each stage of a study, from formulation of the research goals, through data collection and data analysis, to the interpretation of a potential effect of the intervention. In addition, there is ample attention for ensuring the rigor of the research. The presented guidelines are developed and illustrated using an empirical example, in which a constructivist grounded theory approach was applied to evaluate the effect of empathy prompts on the motivation to adhere to COVID-19 regulations.

**Keywords:** Experiment design; Qualitative research; Methodology; Intervention-effect

---

Author contribution: HL & IK developed the theoretical framework for the study. HL & IK designed the study. HL supervised data collection, KS, JT, MB collected data. HL conducted the entire data analysis, JT & MB were involved in the initial stages of data analysis, KS in the later stages of data analysis. HL & IK wrote and revised the article, MB contributed to writing the first concept of the introduction section.

## Introduction

Experimental designs are common research designs in the social and behavioral sciences (Cohen & Lea, 2004; Stangor, 2014). Although there are different types of experimental designs, shortly outlined hereafter, they almost invariably have one thing in common: quantitative data are used to evaluate the effects. In this paper, we demonstrate and discuss the use of qualitative data to evaluate experimental designs, resulting in a methodological tutorial which may be especially beneficial for studies aiming to explain human behavior.

Before elaborating further on the different, qualitative and quantitative approaches to measure the potential effect of an experimental manipulation, we shortly recap the different experimental designs, starting with the distinction between true experiments and quasi-experiments (Gribbons & Herman, 1996). In both true and quasi-experiments participants are subjected to some sort of treatment or manipulation, and the effect of this manipulation is measured on the outcome(s) of interest. In a quasi-experiment participants are naturally members of one of the experimental conditions, without the possibility for the researchers to control the allocation of participants. In true experiments there is full randomization, that is, the participants are randomly assigned to one of the conditions of the experiment. Random assignment enhances comparability between groups as it diminishes the effect of confounders and, therefore, positively affects the internal validity. Although quasi-experiments can lead to undesired differences between conditions, in practice full randomization is not always possible. Throughout the paper, when we use the term ‘experiment’ or ‘experimental design’ this includes both true and quasi-experiments.

A different typology of experiments is the distinction between lab, field, and natural experiments (List, 2007). Laboratory experiments are conducted in highly controlled environments, where researchers manipulate one or more independent variables to create the conditions and subsequently measure the effect of the conditions on the outcome of interest. While laboratory experiments have high internal validity due to the high level of control, they typically have low ecological validity, because the lab situation may be different from daily life (Schmuckler, 2001). Alternatively, field experiments are conducted in the everyday environment while still allowing the researcher to manipulate the experimental conditions. Therefore, in field experiments there is less control over extraneous (confounding) variables, potentially lowering internal validity, but have higher ecological validity (Roe & Just, 2009). Finally, in natural experiments there is no manipulation controlled by the researcher at all, instead

naturally existing groups, that is, conditions, are compared on the outcome of interest in everyday life. Although this can provide valuable, practical information with high ecological validity, there is a higher risk of confounding variables.

Irrespective of the type of experiment the researcher must define the intended outcome, that is, the change as is defined or measured within the study that is assumed to be the consequence of being exposed to an intervention or manipulation. We will refer to the intended outcome throughout this paper as the "*intervention effect*", a term that can refer to quantitative or qualitative outcomes.

In quantitative research the outcome is a measure that reflects the construct intended to be influenced by the intervention. This is generally done by providing operationalizations of the key constructs of interest as quantitatively measurable variables, that are subsequently analyzed by quantitative methods. Using the empirical-analytical paradigm, common in behavioral sciences, one assumes that the universe can be explained through researching empirical data, where a condition for the research is that it is reproducible, replicable, systematic, and transparent (Tijmstra & Boeije, 2011). To achieve these goals a form of reductionism is implied.

Reductionism in research is the process of simplifying a situation to scores on a small number of variables (Verschuren, 2001). Since human behavior is intricate, researchers can engage in using reductionism by bringing complex human beings down to scores on variables. Through reductionism it is feasible to understand if some variables are related, however, quantitative research seldom explains the exact process under investigation. To fully understand human behavior one may need to study the larger context, that is, the complex human being as a whole. Research that focuses on small segments of reality, as is done with quantitative analyses and reductionism, may lack ecological validity by not considering all aspects of the real-life situation and the complex human being of interest.

We therefore argue that it can be beneficial to utilize a more holistic approach for the evaluation of results in experimental designs, especially in the context of behavioral sciences. Holistic approaches are typically associated with an interpretist point of view, and regularly apply qualitative methods (Smaling, 1994). It builds on the idea that human beings and their behavior and motivations are complex and thus need to be studied without reducing them to some numerical variables. Adopting a more holistic view enables researchers to more fully understand the phenomenon of interest and as such can positively influence the ecological validity of a study (Van Steenbergen, 1990). Next to

the holistic approach, qualitative research is often associated with an inductive goal, i.e. theory forming, which can be achieved through an abductive approach (Timmermans & Tavory, 2012). Abductive processes are by nature iterative, as they follow sequences of formulating preliminary explanations based on data which will be tested within the same study with new data. The use of qualitative methods is rather common in several fields of research, and there is much literature available discussing the ways to conduct qualitative research (e.g., Boeije, 2010; Denzin & Lincoln, 2017; Maxwell, 2009; Ritchie et al., 2013), with a clear focus on idiographic knowledge: thoroughly understanding the individual. Since understanding the individual is the focus of much qualitative research, there is little literature on using a qualitative approach specifically for the evaluation of experimental designs in behavioral science.

Some papers offering guidelines for qualitative evaluations of experiments are found in humanities (Naber, 2015), management (Prowse & Camfield, 2013), and political science (Paluck, 2010). To the best of our knowledge, there are no methodological papers in the field of behavioral sciences, but there are a few examples of studies that apply qualitative methods for the evaluation of experiments (Luo et al., 2019; Pelak & Duncan, 2016; Vors et al., 2018), and often they quantify the qualitative results (Rosman & Wilson, 1996; Van Groo-  
tel et al., 2020). These studies offer little to no methodological guidance for other behavioral researchers to apply qualitative methods in their experimental studies.

With this methodological tutorial we aim to provide such guidelines. We provide a step-by-step approach for future researchers aiming for a qualitative evaluation of an experimental design, supported with an example. This worked example describes a field study that investigates the effect of empathy-inducing interventions on the willingness to adhere to social distancing rules during the COVID-19 pandemic. De Ridder et al. (2022) set-up an experimental (quantitative) study for this goal, that provided the framework for another study executed independently from de Ridder (Glebbeek et al. 2024; from here on referred to as GLS). GLS used qualitative methods to evaluate the effect of the intervention following a constructivist grounded theory approach and this study is used as a worked example to develop and illustrate the general methodological guidelines presented in this paper.

The rest of the paper is organized as follows. First, we shortly describe those aspects of the study of De Ridder et al. (2022) that also apply to the qualitative study by GLS, that is the design and intervention, followed by the motivation for the GLS study. In the next section, we present the methodological guidelines

for qualitative evaluation of an experiment, structured in four stages of the research process: defining the research goals, data collection, data analysis, and determining the intervention effect. A separate subsection describes an overarching fifth stage, which pertains the process of reflecting and adjusting during the other stages. The paper concludes with a discussion of the merits of (additional) qualitative evaluation of experimental effects compared to a purely quantitative analysis. Finally, we provide recommendations for researchers also considering a qualitative evaluation of an experiment.

## 4.1 Context of the worked example

In 2020, in the Netherlands, De Ridder and colleagues investigated the effect of an empathy-inducing intervention on the willingness to adhere to the social distancing rules during the COVID-19 pandemic, using a quasi-experiment. One of the measures taken by the Dutch government during the pandemic was social distancing, which included keeping 1.5 meters distance between people.

The field study employed an A-B design (Kratochwill et al., 2010) lasting for 6 weeks, with three sequences of a control week (A) and an experimental week (B) at three campus locations (square outside college hall, main entrance lecture hall, and entrance lecture rooms). The same approach, but with four instead of three sequences of A-B weeks, thus 8 instead of 6 weeks long, was executed a second time four months after the initial wave. In the data-analysis the researchers assumed this to be a between-subjects design, though realistically participants could have been present at multiple occasions throughout the weeks, but without personal identification this could not be detected.

The intervention consisted of prompts to induce empathy and contained three elements: (1) a social robot encouraging people to keep distance; (2) posters of student and staff models with a text expressing a prompt for empathy-based distancing (e.g., 'I have asthma. Keep your distance for me'); and (3) a rail of movie clips of the same models used in (2) with the same texts shown on screens. To determine the effects of the empathy prompts for promoting distancing, De Ridder et al. (2022) used the actual distance between people which was measured using camera recordings. The results of the quantitative study can be found in the paper by De Ridder et al. (2022).

Whereas the camera recordings provided quantitative data on the actual behavior, it does not provide insight in reasons or motivations to (not) adhere to the social distancing rules. To further investigate if and how the intervention had an effect on those aspects, GLS added a qualitative inquiry using the context

of the main study. Certain aspects of the qualitative study were thus set in stone, that is, the intervention and the data collection in the two waves of A-B sequences were predetermined. All other choices were open, providing a good playing field for our study, that is, investigating how to perform a qualitative evaluation of results from experimental designs in the context of behavioral research. While the GLS study is an example of a quasi-experiment, the lessons and guidelines can be applied directly to true experiments as well. Note that the details and empirical results are reported in Glebbeek et al. (2024), while we use the qualitative study as a worked example to derive and illustrate the methodological guidelines presented in this paper.

The studies of De Ridder et al., and GLS as an addendum, were approved by the Ethics Committee of the Faculty of Social and Behavioral Sciences, Utrecht University (file number 20-479). Our methodological tutorial was exempt on the basis of not collecting nor handling data ourselves. We did however have access to the data for research purposes from 2019 through 2023, while the data were anonymized so it could not be brought back to individual participants.

## 4.2 Methodological approach

In this section, we provide a step by step approach developed and illustrated by the GLS study. We explain the methodology followed in GLS, the rationale behind it, and the lessons learned throughout the study.

To ensure transparency and rigor in research, it is common practice to preregister studies and make data open. Also, for the context of qualitative evaluation of experiments these are important considerations before executing the study.

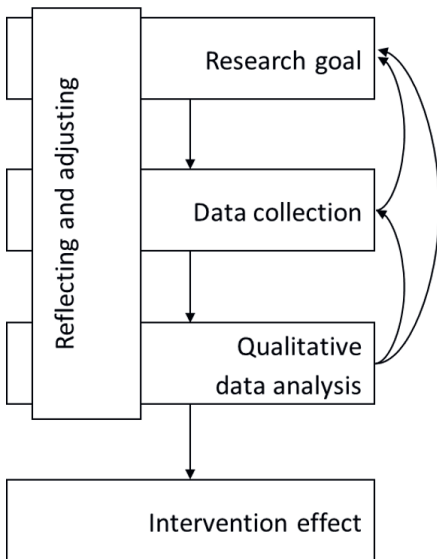
In the context of qualitative studies, preregistration is a debated issue with proponents (e.g. Haven and Van Grootel, 2019) and opponents (e.g. Humphreys et al., 2013). Although the discussion of compatibility of qualitative research and preregistration is beyond the scope of our study, GLS decided to preregister their research to formally document their intended approach before starting the study. The preregistration document can be accessed on the Open Science Framework: <https://archive.org/details/osf-registrations-knb2d-v1>.

Open data is also an essential aspect of scientific research as it enhances transparency, accountability, and reproducibility. However, the application of this practice in qualitative research is also debated (Chauvette et al., 2019; Class et al., 2021). It is argued that openly sharing qualitative data may result in a loss of context and can compromise confidentiality (Chauvette et al., 2019). Despite the impossibility to perfectly interpret the data due to the loss of context, GLS

decided to make the data publicly available through <https://osf.io/e4zfw/> for the purpose of transparency.

After addressing the considerations of preregistration and open data, the actual research can begin. In Fig 4.1, we outline the stages involved in executing a qualitative evaluation of results in experimental designs. While the first stages of defining the research goal, collecting data, and data analysis are common to all empirical studies, the vertical stage labeled 'reflecting and adjusting' is more characteristic of qualitative research and is essential to its success. The arrows pointing upwards on the right-hand side of the figure highlight an essential aspect of qualitative research - the need for reflection and adjustment throughout the research process. It is common in qualitative research to revisit previous stages. This not only improves the quality of the data but also enhances the final analysis and conclusions drawn from the study, including establishing the intervention effect, as shown in the final stage at the bottom of Fig 4.1.

**Figure 4.1: Visualization of the methodological approach of a qualitative experimental study.**



In the following sections, we will provide a detailed description of each stage shown in the figure. We will highlight general considerations that are important at each stage, as well as the specific choices made in GLS.

### 4.2.1 Defining the research goals

In order to conduct a successful qualitative evaluation of an experiment, it is crucial to define its goals explicitly. The open and iterative nature of qualitative research makes clear goal-setting even more critical, as researchers often move back and forth between various stages, which can lead to *scope creep*, that is, a gradual shift away from the original research goals (Roy & Searle, 2020). Therefore, formulating unambiguous goals and regularly reviewing them can help maintain the intended focus, or shift the focus by updating the goals in the desired direction.

Defining clear research goals is particularly important in studies of intervention effects and requires a clear description of the intervention under investigation and the outcome of interest, including how, when, and where the effects are being examined. In some studies intervention effects could be evaluated exclusively through qualitative data: researchers could decide that the qualitative methods provide all the required information to answer the research question. In other studies, intervention effect could also be studied with the qualitative data as an addition to quantitative data within the same study. The latter situation requires that the qualitative component adds value to the quantitative study, for instance, through triangulation or complementing (Bryman, 2006).

In both scenarios, the qualitative investigation of potential intervention effects allows for a more comprehensive understanding of behavior and motivation compared to quantitative research (Van Steenberghe, 1990; Verschuren, 2001). This benefit can however only be achieved with in-depth knowledge of the field.

Therefore, a thorough review of existing literature is required to frame the study and ensure its relevance (Glaser, 1978). Familiarity with leading theories and concepts in the relevant field can lead to the identification of *sensitizing concepts*, which are concepts found in the literature that could be elements of the answer to the research question (Boeije, 2010). Sensitizing concepts can be used to design the study, including data collection and analysis, and can be valuable in defining the research goal. While these concepts are not final, as the definitive concepts and theoretical model will be drawn from the data, sensitizing concepts can guide the study towards achieving its goals.

GLS provides an example of adding a qualitative element to a quantitative study. While the quantitative study focuses on the actual distance between people, which is just one aspect of behavior, the qualitative study aims to provide an understanding of more aspects of behavior and uncover the reasoning behind it. *Empathy* was identified as sensitizing concept and the literature

search showed that *empathy* comprises multiple aspects, these are emotional simulation, perspective-taking, and emotion-regulation (Cuff et al., 2016). Understanding these different facets is crucial for being sensitive to the various ways in which *empathy* can manifest in the study.

The research goal GLS defined for the qualitative study was to achieve a holistic understanding of both the behavior and the motivations for this behavior of participants in the context of COVID-19-related measures. Specifically, they aimed to identify the strategies that participants employed to adhere to the COVID-19-related measures and the motivations behind their choices. This allowed them to test the underlying theory on the effectiveness of prompts, with a predominantly inductive approach instead of a more narrow assessment of the behavioral outcomes.

### 4.2.2 Data collection

In qualitative research, just as in other studies, data collection needs to be carefully planned to ensure that the chosen methods generate data that can effectively address the research question. It is both advisable and considered a standard practice in qualitative research to initiate data analysis as soon as part of the data becomes accessible. This proactive approach not only encourages critical reflection on the research process but also facilitates timely adjustments to data collection methods when needed. Typical data collection methods in qualitative research are observations, interviews, focus groups, and archival data (Carr et al., 2021). It is important to note here that in the context of an experiment, the content of the intervention might limit the options for plausible data collection methods. Perhaps the most common methods, and what GLS used in their study, are observations and interviews. Therefore, we will focus on discussing the strengths and limitations of these two methods.

Observations are particularly valuable as they enable the study of naturalistic behavior and covert field observations can achieve a high level of ecological validity (e.g., Angrosino and Rosenberg, 1994; Mulhall, 2003). Two types of observational studies are distinguished: systematic and non-systematic. Systematic observation involves using a pre-defined checklist to quickly and accurately record behavior with high inter-rater reliability (Reiss, 1971). However, it may not offer the holistic understanding, since what is being observed is determined beforehand. In contrast, non-systematic observation allow researchers to capture all relevant aspects of the situation in their field notes, which provides a detailed description of the behavior and its contextual setting. This approach

can result in a more thorough understanding of the behavior and the factors that contribute to it. However, non-systematic observations are more prone to observer effects, that is, lower inter-rater reliability (Mulhall, 2003).

Interviews are a valuable research tool for gathering detailed information on the experiences, rationales, or attitudes of individual people (e.g., Dearnly, 2005; Evers, 2015). While interviews may not provide information on naturalistic behavior, they can offer insight into intentions and experiences. Two main types of interviews are distinguished: open and semi-structured. In open interviews only a starting topic is selected and the interviewer and interviewee discuss what seems relevant to them. Open interviews can provide a holistic understanding of the individual, allowing the interviewee to address topics the researchers may not have considered. However, the lack of structure in open interviews can lead to some topics being overlooked, creating differences between interviews, which is an example of an interviewer effect that may limit the quality of data obtained (Evers, 2015).

Semi-structured interviews on the other hand follow a more predictable pattern by pre-determining to a greater extent the content of the interview, which enhances comparability across interviews but may limit exploration (Evers, 2015). To structure such interviews and ensure comparability, it is recommended to use a topic list that includes the topics to be discussed, their order, and specific wording of questions (Kallio et al., 2016). Such a list can help focus on relevant information and reduce unwanted researcher bias (McGrath et al., 2019). It also enables researchers to ask relevant questions that are more likely to elicit valuable data on the relevant concepts.

In qualitative research, one common practice is to employ method triangulation, which involves using multiple types of data within a single study (Korstjens & Moser, 2018). By using method triangulation, researchers can study the phenomenon of interest from multiple perspectives, which ultimately enhances their understanding of it. Method triangulation allows researchers to choose data collection methods that complement each other's strengths and weaknesses. For instance, while interviews may provide insight into intentions, observations may provide information on actual behavior.

In GLS, the researchers utilized both systematic and non-systematic observation methods to investigate whether individuals adjust their behavior in response to prompts and to identify circumstances in which individuals adhere to or violate COVID-19-related measures. For systematic observations, a simple checklist was provided to record whether individuals seemed to notice the prompts, kept their distance from others, and if they walked alone or in a group

(recording group size). Non-systematic observations were included through field notes, that provided a more detailed description of individual behavior, their relationship with potential companions, and changes in behavior related to adherence to COVID-19-related measures. Participants were informed about data collection through signs with explicit information about the research. In addition the project's purpose and procedures were thoroughly explained on the university website.

Six researchers conducted these observations over a seven-week period in the first wave of data collection. Assigning the same task to multiple researchers is known as researcher triangulation and will reduce the influence of individual preconceptions (Archibald, 2016; Patton, 1999). A pilot week was used to determine the observation schedule and test the checklist, followed by three A-B iterations. The A-weeks served as the control condition with no intervention, while the B-weeks were the experimental condition. Topic lists for both weeks are included in Appendix 5.A<sup>1</sup> and Appendix 5.B. A total of 52 hours of observations were conducted for the control condition, and 70 hours for the experimental condition, observing a total of 1002 individuals. The sample consists predominantly of highly educated adolescences, since the study location is almost exclusively visited by students and employees of the university and the university of applied sciences.

In addition, GLS used method triangulation through short interviews that complemented the information from the observations. The same six researchers conducted 5-minute interviews while walking with the people they observed, in the last week of data collection of the first wave. The researchers used a topic list to guide the conversation aiming for information about the motivation for their behavior, added in Appendix 5.C. A total of 42 participants were interviewed, equally split between the control and experimental group. The interviews were transcribed intelligent verbatim to ensure the preservation of meaning (McMullin, 2023).

The second wave of data collection was announced at a later stage and took place half a year after the first wave. The duration of the second wave of the study was eight weeks in total and consisted of four sequences of a control and an intervention week. The intervention was the same as in the first wave and was also placed at the same locations. The second wave enabled GLS to collect the data needed to answer the research question and the design was informed by the results from the first wave since data analysis was already conducted for that part of the data. Both the short and the long interviews were preceded

<sup>1</sup> All Appendices for this chapter are included at the end of Chapter 5.

by asking respondents for consent to record the conversation and to use the collected data for scientific publication. Their consent was also audio-recorded.

The key objective for the qualitative data collection in this wave was to gain a better understanding of the motivations and strategies of participants. To obtain more in-depth data, the researchers conducted longer interviews (approximately 45 minutes) with a predefined topic list to enable a comprehensive understanding of the processes that potentially influenced the participants' behavior. The topic list was based on the data collection of the first wave and contained topics like *Social life*, *Home situation*, and *Study-related contact*. For each topic, a starting question was predefined (e.g., 'Can you describe your recent interactions with fellow students and lecturers?') as well as some follow-up questions (e.g., 'How does this differ from before corona?') and subtopics (e.g., *agreements*, *activities*, *group composition*). The topic list is added in Appendix 5.D.

The interviews in the second wave of data collection were done by five researchers at one of three locations: a building near the interventions, another university building away from the interventions, and online. Of the 39 participants, six were interviewed in the intervention-weeks and near the intervention location and were therefore considered part of the experimental group. The control group consisted of 33 participants who were interviewed online (3), on a location away from the intervention (13), or near an intervention location but during a control week (17). The participants of both types of interviews were all students at the university or the university of applied sciences. Besides three non-Dutch participants in the control group, all participants were Dutch natives. While participants could have been the subject in multiple observations, no participants were interviewed more than once. The interviews were transcribed intelligent verbatim in order to stay true to the words of the respondent.

By using non-systematic and systematic observations as well as short interviews and semi-structured interviews, GLS aimed for a comprehensive overview of the potential effect of the intervention. By combining data from these different sources, they expected to be able to describe the relation between behavior and underlying motivations, attitudes and intentions with respect to COVID-19-related measures and compare results between groups with and without exposure to the intervention.

As stated earlier, it is common practice to start initial data analyses simultaneously with collecting data in qualitative research. This can offer valuable insights and enables researchers to improve their data collection methods or

identify confounders not anticipated beforehand. Before we discuss the qualitative data analysis in the next section, we first elaborate on the process of reflecting and adjusting that is typical in qualitative research and is needed to ensure the quality of both the collection and the analysis of data.

### 4.2.3 Reflecting and adjusting

Maintaining rigor throughout the research process is critical to ensuring the overall quality of the research. Four critical aspects of rigor include dependability, confirmability, transferability, and credibility (Guba, 1981; Korstjens & Moser, 2018). Dependability and confirmability emphasize the importance of minimizing the researcher's influence (Lincoln & Guba, 1985). Rigorous studies should produce objective findings that are not influenced by the researcher's preconceptions (*confirmability*) and are consistent across different contexts and time frames (*dependability*). Credibility and transferability focus more on the quality of the findings within their specific context (Lincoln & Guba, 1985). Rigorous results accurately reflect the truth value and the participants' perspectives on the phenomenon of interest (*credibility*) and are applicable in a wide range of situations (*transferability*).

Qualitative research can ensure rigor through various methods, including applying thorough data collection methods, data triangulation, researcher triangulation, and reflexivity. While the first three have been introduced earlier in this paper, reflexivity is a term that has not been discussed yet.

Reflexivity can target diverse facets of the research process. In order to address all pertinent dimensions, we categorize it into two distinct forms: personal and epistemological reflexivity. Personal reflexivity addresses the unintended influences of the researcher's values, beliefs, social identity, and experiences on the study (Palaganas et al., 2017). It also acknowledges that researchers are active participants in the study, and their biases can shape the way they perceive and approach the research world. For instance, researchers may be more likely to note down observations that align with their expectations, and overlook or disregard unexpected findings. Being aware of these preconceptions can help researchers identify their unwanted influences on the data and findings.

Epistemological reflexivity on the other hand involves an examination of the potential impact of the study's design on the research outcomes (Mauthner & Doucet, 2003). Epistemological reflexivity enables researchers to reflect on questions such as, "Could alternative methods be employed to investigate the research questions?" or "How does the study's design impact the data and find-

ings?”. Engaging in epistemological reflexivity allows researchers to critically evaluate the data’s origin and how the study’s design has influenced specific findings. Data collection can impact the direction of research, and epistemological reflexivity can help researchers recognize this and make necessary adjustments.

Saturation is another critical term in qualitative research, referring to the point where new data does not yield any new information (Fusch & Ness, 2015). Achieving saturation is often a goal for researchers as it allows them to determine when data collection can stop. To achieve saturation researchers employ theoretical sampling, where participants are selected based on their potential to advance the research (Coyne, 1997). This approach requires analyzing data during data collection to determine the characteristics of subjects that can further the research. Researchers can focus on negative cases, those that differ significantly from the majority of cases, to ensure the study’s completeness (Henry, 2015).

At the beginning of a qualitative study, it is often unclear what constitutes sufficient data, that is, enough data to achieve saturation. Also the research focus may shift throughout the study. Therefore, it is recommended to commence data analysis as soon as part of the data is available and continuously reflect on the research process, making adjustments as necessary, as this is inherent in qualitative studies.

GLS employed multiple measures to ensure the rigor and quality of their research. Firstly, they utilized method triangulation by conducting both observations and interviews to increase the credibility and reliability of their findings. Additionally, they employed researcher triangulation during data collection and analysis to reduce the potential for bias, that is, confirmability, and increase objectivity. They also used theoretical embedding to facilitate the interpretation of the results by incorporating sensitizing concepts.

To achieve data saturation within the predetermined fixed time available for data collection, GLS collected as much data as possible. They used theoretical sampling to increase the efficiency of data collection by providing their researchers with specific characteristics to look for, such as non-romantic couples or larger groups walking together, and by paying attention to negative cases. Initially, GLS recorded participants’ gender and age during the interviews. However, it soon became evident that these demographic characteristics were not analytically relevant to the aims of the study, as no meaningful patterns related to demographics emerged from the data. Instead, the social contexts in which participants operated proved to be of greater relevance. Consequently, GLS discontinued the systematic documentation of demographic variables and

redirected their focus toward aspects more pertinent to the research objectives.

They also incorporated various methods to promote both forms of reflexivity in their study. They held regular meetings with different groups of researchers to discuss the collected data and preliminary analyses. Additionally, they conducted weekly meetings with researchers to monitor the progress of data collection and review the field notes. Through this process, GLS identified that the field notes lacked sufficient detail on certain aspects that the observers took for granted, such as the relationships between people walking together. To address this issue, they instructed the observers to take more detailed factual notes to enrich the data.

Finally, through epistemological reflexivity, GLS recognized that the data collected during the first wave was insufficient to answer their research question thoroughly. Rather than adjusting the research goal, they decided to use the second wave of data collection to gather the additional data required, hence the decision to use semi-structured interviews, with the topic list representing the relevant themes to discuss.

#### 4.2.4 Qualitative data analysis

Throughout data collection, the (initial) analysis of data and constant reflecting and adjusting play a pivotal role. Also within data analysis, the stages presented in this section are subject to constant reflection and potential going back and forth between them. Nevertheless, we structured this section in three parts, describing subsequent stages, to enhance clarity. In the first subsection, the preparation phase is discussed. Initial choices concern how data are formatted, what software is used, and which analysis framework will be applied. We will emphasize the role and impact of such choices specifically for the context of experimental qualitative data analysis.

Given the choice for a constructivist grounded theory approach, the actual analysis contained the phases of open and axial coding (second subsection) and the final phase of selective coding (last subsection). Selective coding results in the construction of a theoretical model. GLS initially only had data from the first wave and processed and analyzed it up until the selective coding phase. With the realization that information was lacking for the goals at hand and with the opportunity of a second wave of data collection, the final analysis used data from both waves. So, although the second wave data was needed to be able to construct the final theoretical model, it is important to note that all data undergo rigorous processing throughout all stages of data analysis. Illustrations

of this process will be provided throughout this section.

### **Preparations**

Some considerations and choices need to be made before starting the actual data analysis. Not only are there different analytical approaches and several software packages for qualitative data analysis, there are also different ways to make group comparisons when using qualitative data to evaluate effects in experimental designs. An important aspect in making some of these choices is the outcome that the study at hand aims for, as determined by the type of research question that was formulated. Kearney (2001) makes a distinction between descriptive, exploratory, and explanatory research questions. Descriptive and exploratory questions aim for a relatively superficial understanding of aspects or themes at play in the situation being studied and typically result in models that remain close to the data. In contrast, explanatory research aims to provide a more thorough understanding of the process under investigation and often leads to the development of more abstract models, for which the data is transformed to a higher degree (Sandelowski & Barroso, 2003). An experiment is to some extent inherently explanatory, since the main goal is to evaluate if an intervention has effect on an outcome. However, the outcomes of interest can be more or less precisely defined, *a priori*, and this determines whether the study is more exploratory, descriptive, or explanatory.

An often used paradigm in qualitative studies is constructivist grounded theory (CGT) as developed by Charmaz (2017). The CGT approach aims to represent the perspectives of the participants and therefore fits well with behavioral research. The data is processed in a structured, step-by-step approach in such a way that the data speaks for itself. Some qualitative researchers argue that the high level of structure of CGT limits the creative space characteristic of qualitative research (Henwood & Pidgeon, 1995). Other approaches, such as discourse analysis, narrative analysis, and phenomenology, are less structured than CGT and can be more interpretative. Discourse analysis seeks to explain behavior from an outsider perspective (Hodges et al., 2008), narrative analysis focuses on reconstructing complete stories (Franzosi, 1998), and phenomenology aims to describe experiences (Wojnar & Swanson, 2007). Depending on the context of the research each of these approaches could be used for the evaluation of experiments.

The treatment or intervention of interest creates groups in the data. In a qualitative analysis of an intervention effect, one can incorporate this grouping in two ways. First, all data can be analyzed as if coming from one population

(e.g. Palsola et al., 2020; Pelak & Duncan, 2016). In this case, the last phase of the analysis is comparing the resulting model between the groups. The description of a potential intervention effect is in terms of certain parts of the model originating from one group and not the other, or in terms of certain relations being more prominent in one of the groups. This approach may be the preferred choice when the research question is merely descriptive or exploratory; it may reveal unexpected findings, such as themes that are entirely missing from one subgroup or a process that is not part of the other. Alternatively, separate analyses can be done for the treatment conditions (e.g. Prowse & Camfield, 2013). Then distinct models are built for the data from each of the groups and the final phase consists of comparing the resulting models. This approach is suitable when it is assumed that groups differ inherently and independent processes are expected to be evaluated. This approach seems more natural for studies where aims and questions are of a more explanatory nature.

Irrespective of the analytical approach and the way the grouping structure is incorporated, appropriate software is required to systematically analyze the data. Various computer-assisted qualitative data analysis software (CAQDAS) programs are available and assist researchers in structuring their analyses by identifying (sub)groups in the data, structuring codes, creating separate models, and running queries. Well-known CAQDAS programs include MAXQDA (2021), Atlas.ti (2022), and NVivo (2020), among others. All these programs provide functionalities such as coding, compatibility with multiple data types, specification of demographics per document or case, memo writing and linking, query usage, and project file sharing. Researchers can choose a CAQDAS based on their specific needs, familiarity with the software, or availability.

GLS used CGT to answer the explanatory research question on how the empathy-inducing nudges of the intervention program potentially influenced the decision making process with respect to adherence to the distance regulations. NVivo Release 1 (2020) was used for the data analysis, since it satisfied their criteria and was readily available through a campus license. Within NVivo, they distinguished between the control and intervention groups by creating attributes for each case, including control or experimental group, day, time, interviewer/observer, and week.

Despite the explanatory nature of the research question, GLS assumed that similar processes would be present in the control and experimental conditions. They therefore choose to analyse all data together to create one overall model. However, the intervention could lead to different weights in parts of this model, for instance, certain concepts or relations may be only present in one of the

groups, while other parts of the model may be more prominent in one group than in the other. If such group differences emerge, GLS can conclude that the intervention imposing empathy-inducing nudges triggered different processes and thus was effective.

### **Open and axial coding**

In the remainder of this section we will describe the actual data analysis given the choice for a CGT approach. In the grounded theory tradition, three phases in the coding of data are subsequently conducted: open coding, axial coding and selective coding (Corbin & Strauss, 1990; Glaser, 1978).

During open coding, the aim is to identify and list all topics presented in the data by allowing to let the data speak for itself. Preconceived notions may interfere with the analysis and, therefore, it is important to avoid imposing structure at this point. To ensure the quality of open codes, researchers should keep an open mind and be prepared for unexpected themes to emerge. However, to fully understand the phenomenon of interest, it can also be beneficial to incorporate the earlier identified sensitizing concepts in this phase of coding. Although the sensitizing concepts are typically too broad to include directly, they can be translated into so-called a priori codes (Stemler, 2000). A priori codes are more specific than sensitizing concepts, and provide a link between data and theory. The use of a priori codes enables identification of the specific ways in which sensitizing concepts might occur in the data and helps interpret the data. An example of a sensitizing concept could be *empathy*, which is a theoretical, abstract construct. Participants will most likely not mention or express this construct directly, so a priori codes are needed. Examples of a priori codes for *empathy* could be 'protecting a loved one' (for interview data), or 'steps back so other people can join a conversation' (for observational data).

Effective open coding involves adhering to the 3C's: content, context, and coverage. This entails applying labels that describe the content of a fragment, which spans enough data to capture the context, while ensuring that the created codes cover all topics in the data (Williams & Moser, 2019). The open coding phase continues until saturation is reached, that is, until adding new data does not lead to new open codes (Glaser & Strauss, 1967).

The next phase is axial coding, which aims to structure the data and identify latent themes by exploring the relationships between the open codes that emerged from the data (Scott & Medaugh, 2017). Although sensitizing concepts can be used to provide an initial structure, it is important to remain open-minded and avoid tunnel vision. By exploring different possible struc-

tures, researchers can gain a more nuanced understanding of the various layers in the data and potentially uncover new perspectives that were not previously considered. Throughout this process, the method of constant comparison is applied to ensure that themes are consistent across the data (Giles et al., 2016). This involves checking how themes are present in the data of other participants and how they may differ in terms of their characteristics.

Axial coding results in a structured code tree, with well-defined main codes (themes), and relevant sub-themes and open codes placed underneath the overarching themes. Before discussing the next phase in the data analysis, we will first continue with a description of the initial analyses in GLS, and present some additional considerations that are important before moving to the third phase of analysis: selective coding.

In GLS, the open coding of the data from the first wave was conducted by six researchers ensuring researcher triangulation. Although GLS considered creating a priori codes for the sensitizing concept *empathy*, they decided against it to avoid over-fitting the data and over-interpreting the findings. Codes were created to describe the content, selecting fragments to capture their context, and cover all themes in the data as prescribed by the 3C's approach. For instance, they produced codes such as *Waits for passersby to pass*, *Avoid putting entire house in quarantine*, *Keeping your distance with friend is weird*, and *Avoids grandparents due to health issues*. Saturation was achieved after coding approximately 60% of the observations and 80% of the short interviews.

The axial coding phase was carried out by two of the six researchers doing the open coding, and two additional researchers. In this phase the sensitizing concept *empathy* was applied as one of the themes, since they wanted to investigate its impact on adhering to COVID-19-related measures. Data from the observations served well for the identification of behavior and behavior changes. The data from the short interviews was supposed to add insights into participants' strategies and motivations, but in this phase of analyzing the data it became clear that the short interviews provided insufficient information to fully address their research goals. With the current data GLS could only provide a description of visible behavior, the direct influence of the nudges, and a very general enumeration of considerations for the behavior. It became clear that an in-depth understanding to explain which motivations were active in both groups, and how that led to certain strategies to behave in relation to COVID-19-measures, would not be achieved.

In such a situation there are two options: collecting additional data (if possible), or reporting the insights gained from the available data, even if the original

questions cannot be fully answered. GLS were fortunate that a second wave of data collection was announced, allowing them to collect more in-depth data and to pursue a complete answer to their original research question.

Resuming the data analysis with the additional data from the interviews, in the axial coding phase they initially still used the overarching theme of *empathy* to view the data through this lens. With the data from the second wave however, at some point they decided to drop *empathy* as a sensitizing concept. It appeared to lead to over-fitting the data, since most themes could be interpreted as sub-themes of *empathy*. Excluding *empathy* enabled the researchers to gain a more comprehensive understanding of the data.

The new data provided more context and explanation to the observations' results and supplemented the information acquired from the short interviews. To address the research goal, two main themes were formulated: *Motives* and *Strategies*. Several sub-themes were identified that represented aspects of these main themes, e.g., 'no fear' under motives, or 'use of self-tests' under strategies. A further illustration of part of the coding scheme is provided in Table 4.1 and shows how themes and underlying sub-themes emerged from observations and text fragments from the interviews. Such a visualization of results forms the basis for the final phase, that is, selective coding.

### Selective coding

Selective coding is the third and final phase of data analysis in the grounded theory tradition (Strauss & Corbin, 1990). This phase is the culmination of the study process, which allows the qualitative researcher to engage in a dialogue with the data to search for the most accurate representation of the truth value. Its main objective is to develop a theoretical model that explains the phenomenon under investigation.

The outcome of axial coding, which includes the code tree, definitions, and memos generated during the open and axial coding phases, serves as the foundation for building the theoretical model in the selective coding phase. The selective coding process involves identifying the relationships between the themes that emerged from the previous phases, such as developing a chronological story, typology, or identifying which themes can or cannot coexist (Boeije, 2010). Providing clear instructions for this phase can be challenging, as the approach may vary depending on the research situation. A common strategy is to identify a core category that connects the themes previously identified (Giles et al., 2016). This strategy is particularly useful when multiple models are created and each sub-population needs to be analyzed, as is typically the case in experimental

**Table 4.1: A selection of subthemes, open codes, and fragments to serve as illustrations for the coding scheme resulting from the axial coding phase.**

| Theme                              | Subtheme             | Open code                                 | Fragment                                                                                                                                                                                                                                                                            |
|------------------------------------|----------------------|-------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Motives                            | Influence others     | Mother sees little risk in being close    | ‘Well my mother, she works at home. And she only goes to the supermarket and back. There is so little risk of being infected.’<br>‘That is how I rationalize it: she is vaccinated, she sees barely any people, I don’t see many either. We think it is okay.’                      |
|                                    |                      | Keeping friends at a distance feels weird | ‘It is so weird to keep your distance if you are not used to doing that. That is why we don’t do that.’<br>‘If I wanted to [edit: keep distance between friends], it is not realistic.’                                                                                             |
|                                    |                      | ”Hugger” kept giving hugs                 | ‘One friend was always used to giving hugs. And he kept doing that. But with other friends we give a box.’                                                                                                                                                                          |
|                                    | No fear for COVID-19 | Risk of getting infected is low           | ‘For as far as I can tell, the risk for us of getting infected is low.’<br>‘My friends, the four I regularly see, are very calm, timid, boys. They don’t go out or do crazy student things. So if you are at home, the chance is much lower of getting infected. I take that risk.’ |
|                                    |                      | Risk of infection is low                  | ‘I am still young, I am not afraid I will get really sick from the virus.’                                                                                                                                                                                                          |
|                                    | Strategies           | Use of self-tests                         | Tests regularly                                                                                                                                                                                                                                                                     |
| Testing before meeting family      |                      |                                           | ‘Before going home for Christmas I did a self-test to be sure that I had not caught COVID-19, so I could go home without too much risk.’                                                                                                                                            |
| Limited number of people gathering |                      | Mostly one-on-one meetings                | ‘I meet friends mostly one on one.’<br>‘We had the rule to only allow one external person in the kitchen at any time.’                                                                                                                                                              |
|                                    |                      | Avoiding large gatherings                 | ‘We are relatively safe, we do not organize or attend any large illegal parties.’<br>‘Well I invite less people to my home.’                                                                                                                                                        |

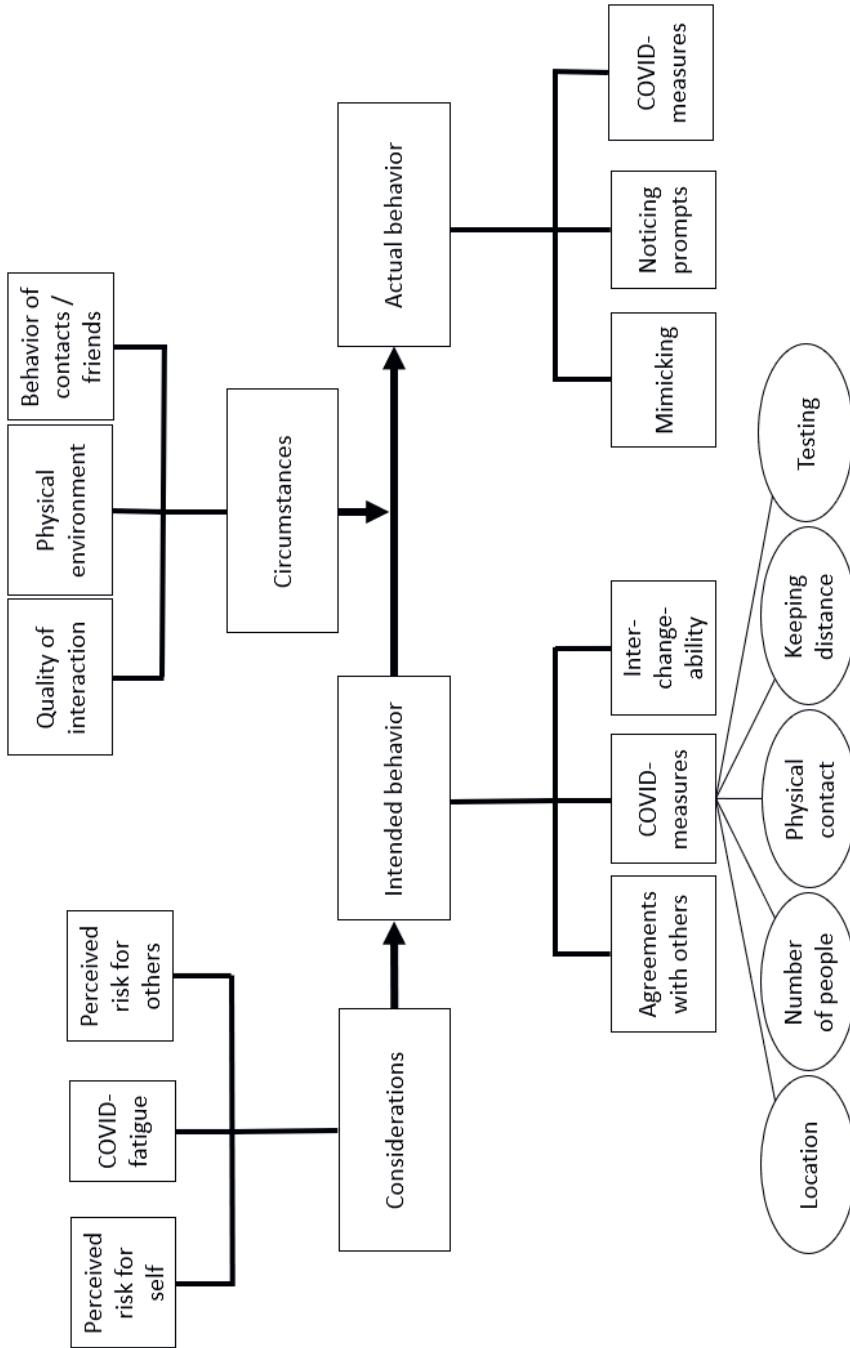
designs. The core category acts as the missing piece that connects the different models and explains differences between the sub-populations. This will be further elaborated in the subsection about evaluating the intervention effect.

Throughout selective coding the process of *constant comparison* is applied (Boeije, 2010). With constant comparison researchers continuously check how modifications to the model work for other, already analyzed, data. A potential pitfall to qualitative research is to let the final round of data collection be the deciding factor for the final model, that is, the patterns present in the last observed data are used to determine the definitive descriptions. Through constant comparison, each addition or adaptation of the model is verified with earlier data. This process is essential to ensure that the final model correctly represents the participants' views.

The outcome of selective coding is a theoretical model that explains how participants perceive the phenomenon or process of interest (Sandelowski & Barroso, 2003). The resulting theoretical model should adhere to four key principles: they should *fit* within the field of study, be *understandable* to those in the field, be *generalizable*, and provide some level of *control* to the researcher in the form of understanding the process or phenomenon of interest (Glaser & Strauss, 1967). By setting up the study as described in the previous sections, the results acquired should meet these criteria. The use of sensitizing concepts helps fit the theory within the existing literature, while theoretical sampling ensures generalizability. The CGT coding procedure ensures that the resulting theory is understandable and provides control to the researcher.

To illustrate the process of creating and verifying a theoretical model from the qualitative data, we present the results of GLS in Fig 4.2. Here, we zoom in on a few aspects of the model, to give an impression of the process and result of selective coding. Note that in this phase they did not distinguish between participants in the intervention or control condition; they built one overall model for all data obtained.

Figure 4.2: Theoretical model as a result from the selective coding, where forms of behavior are explained by different kinds of circumstances, intended behavior, and considerations.



In GLS, the stories of participants demonstrated some common ground that helped understand different aspects of their behavior and motivations. For instance, the researchers observed a recurring pattern in the participants' explanation of circumstances that affected their choices for actual behavior. From observations and interviews it became clear that social interactions are context-dependent, that is, behavior does not only depend on intentions (based on reasoning about measures and the motivations to adhere to them) but also on, for instance, how friends or other contacts behave, what their intentions are, and on perceived risk of not behaving in the intended way (see Fig 4.2). To give just one illustration, the following quote is exemplary for several similar remarks about the influence of other people's behavior:

Participant D2505B2: *'It is sort of legitimized because the people around you don't do it either.'*

To verify the conclusion that social interactions are context-dependent, a revisit of all data was made to validate this pattern with the data itself as the method of constant comparison describes. An example of such a verification in the data was given by a participant that described intended behavior depending on who they were with.

Participant D1006O1: *'When I'm with family, then I'm a bit easier about it [edit: 'it' refers to keeping distance]. But if I'm just in the supermarket then I'm a bit more careful with that. So that's actually kind of where I draw the line. People close to me, that's where I am very lenient with it. But if it's just about public places then I am more cautious.'*

The researchers saw similar things in the observations happening. Most people ignored the station for hand sanitation, but the following happened when people entered together:

Observation M1611D215: *'PP does not keep distance from either friends. First friend disinfects their hands, they stand close to each other while waiting. When the first friend is done, PP disinfects their hands, and waits until the third person has disinfects their hands as well. When they walk away they do not keep distance.'*

Similar things happened on a regular basis: when one person of a group adheres to a measure, the others follow their lead. With the developed theoretical model, GLS could more easily classify new data and categorize topics mentioned by the participants.

### 4.2.5 The intervention effect

The final phase in determining evidence for a potential intervention effect is the comparison of the intervention and control group. In research where separate models are created for the groups, this means comparing the models on their similarities and differences. This entails describing the processes per group, in order to understand at what point they take meaningfully different paths. Subsequently, the data is again consulted to get a comprehensive understanding of participant's perspectives in both groups. The final challenge is to describe them accordingly, so the readers can grasp the essential characteristics for both groups.

In research where a singular model is employed, researchers often utilize queries as valuable tools provided by CAQDAS. These queries generate cross-tabulations, which elucidate the frequency of various aspects of the theoretical framework observed within different groups. A basic query involves tallying the occurrences of specific (sub)themes across these groups. Furthermore, researchers can analyze the co-occurrence of themes and discern potential disparities between intervention and control groups regarding the prevalence of theme relationships. However, it is imperative to note that while queries offer initial insights into the underlying processes, a thorough understanding and interpretation of the results within a theoretical framework necessitate continual data analysis through constant comparison.

In addition to quantifying the results with queries, a qualitative assessment is indispensable. Qualitative researchers often describe this process as "engaging in a dialogue with the data". Essentially, researchers aim to uncover potential relationships using various tools such as queries, memos, and theoretical frameworks. These relationships are then rigorously tested by revisiting the original raw data, a practice akin to the concept of constant comparison introduced earlier.

GLS categorized all data entries as belonging to either the intervention or control group. Table 4.2 provides an illustration of some comparison between these groups. The researchers chose the themes to inspect partly based on the theory behind how prompting with empathy would influence behavior: specifically the subthemes under *motivation*. The table displays the proportions of participants from each group who reported a particular theme or combination of themes. By examining these proportions, GLS could gain initial insights into the potential differences between the groups and identify the particular aspects in which they may diverge.

**Table 4.2: Example of query used to process selective coding to crosstabs: proportion of participants mentioning subthemes of strategy, motivation, and relations between subthemes in the control and intervention groups.**

| Used strategy                                 | Control | Intervention |
|-----------------------------------------------|---------|--------------|
| Limited number of people                      | .64     | .33          |
| Meeting outdoors                              | .24     | .33          |
| Use of self-tests                             | .15     | .50          |
| Behavior contacts/friends                     | .64     | .67          |
| Motivation                                    | Control | Intervention |
| No fear for COVID-19                          | .16     | .50          |
| Perceived risk for self                       | .64     | .67          |
| Duration of measurements                      | .36     | .67          |
| Relations                                     | Control | Intervention |
| Perceived risk for self vs. Use of self-tests | .12     | .33          |
| Perceived risk for self vs. Meeting outdoors  | .12     | .00          |

Subsequently, the investigated queries prompted the generation of preliminary explanations, which in the context of abductive reasoning (Timmermans & Tavory, 2012) can be seen as a form of (data-generated) hypotheses. To illustrate the analytical process, let's consider the theme of *No fear for COVID-19*, which was mentioned less frequently in the control group compared to the intervention group (control: .16; intervention: .50). The researchers' initial expectation was that this difference might have resulted in divergent intentions to adhere to COVID-19-related measures. However, upon revisiting the data, no clear differences in intentions between the groups were observed, that is, the intervention group did not overall seem less willing to adhere to COVID-19-measures than the control group.

To explain the finding that the intervention group, despite mentioning the motivation *No fear for COVID-19* more frequently, expressed similar intentions to adhere to COVID-19-related measures, a new preliminary explanation was formulated. GLS speculated that differences in *empathy*, which the intervention aimed to induce, could account for the absence of divergent intentions. It was theorized that although fewer participants experienced fear for COVID-19 themselves, the empathy-based nudges might have influenced their intended behavior, resulting in a comparable level of adherence to measures. However, upon conducting another reassessment of the data, this preliminary explanation was not supported either. Instead, during this round of data reassessment, two other topics consistently emerged in conjunction with the mention of *No fear for COVID-19*. It became apparent that participants who expressed no fear for

COVID-19 often mentioned employing specific strategies to mitigate the risk, such as using self-tests and limiting their social interactions to a small number of people. The following quote serves as an illustration of this observation:

Participant E1905L1: *‘I have one group of friends that I actually see [edit: without keeping distance] every week. Yes, if they all stay within that group of friends, then it’s not too bad with how many contacts you have’.*

4

The researchers’ final theory was that, notwithstanding minor differences between the groups, no relevant effects attributable to the prompts were observable. To validate this assertion, they conducted a test wherein two researchers were tasked with reviewing two transcripts of semi-structured interviews, while the experimental status was masked, and asked to allocate these to either the control or experimental group. The researchers were unable to accurately allocate the transcripts, supporting the preliminary explanation that the prompts had no relevant effects in this study. A further discussion of the empirical findings of the COVID-19-study is beyond the scope of this paper and can be found in Glebbeek et al. (2024).

With this the data analysis is concluded. Our intention was to demonstrate the iterative process of preliminary explanation generation and validation through repeated data reassessment. This iterative approach is vital in answering the central question of qualitative evaluation: how can we describe and comprehend the effects of the intervention on the desired outcomes? By engaging in this meticulous and iterative process, researchers strive to gain a comprehensive understanding of the intervention’s impact on participants’ behavior and the motivation for this behavior. In the context of the quantitative (De Ridder et al., 2022) and qualitative (Glebbeek et al., 2024) studies referred to in this paper, in both studies it is concluded that no support for an intervention effect was found. However, by developing a model based on empirical qualitative data, GLS were able to provide deeper insight into the underlying mechanisms behind the quantitative findings, offering an explanation of the internal processes that contributed to those outcomes.

## 4.3 Discussion

In this methodological tutorial, we provided guidance to assist (quantitative) researchers in incorporating qualitative methods into their experiments. Compared to existing literature in this area, we aimed to provide a practical and comprehensive overview of the stages of the research process from start to finish. Furthermore, the focus on the application of qualitative research to experimental designs is a contribution in the field of behavioral sciences that was missing until now. We argued that, by following the stages outlined in Fig 1, researchers can obtain a more holistic understanding of their research problem than with the use of solely quantitative methods. Here, we add Table 4.3, which summarizes the essential elements for conducting a rigorous qualitative investigation of potential intervention effects in experimental designs, and gives an overview of some key references for these elements. With this paper we aimed to guide researchers who are well-versed in quantitative methods towards the integration of qualitative methods into their study. By providing detailed guidance and an illustration, we hope to facilitate a smooth transition and ensure the quality and rigor of qualitative experiments.

Previous literature on qualitative experiments in behavioral sciences primarily engaged in philosophical contemplation regarding the suitability of qualitative methods within experimental contexts (Naber, 2015; Paluck, 2010), or provided individual empirical reports (e.g., Luo et al., 2019; Pelak & Duncan, 2016). In contrast, our focus is on highlighting the necessary stages that researchers must undertake to conduct their own qualitative evaluation of an experimental intervention. Where earlier studies focused mainly on the design of the research (Prowse & Camfield, 2013), we discuss considerations and decisions for all stages of the project. Notably, our contribution places a greater emphasis on the application of reflexivity throughout the study and advocates for a flexible and iterative process.

The degree of flexibility inherent to qualitative research is one of the main differences compared to the quantitative approach. Refining the study design throughout the research process adds to the rigor of the research. In quantitative research, there should be a clear a priori plan for data collection and analysis. If it is needed to explore how different elements work, a pilot study can be employed before the actual study, and these pilot data should not be included in the actual analysis. In contrast, in qualitative research adjustments to the study design and analysis can be made as needed throughout the study. Reflexivity should improve the rigor of the study and assist in finding the most reliable and

**Table 4.3: The most relevant topics to consider when evaluating experiments with qualitative methods, ordered per stage of the research.**

| Stage                  | Short description of stage                                                                                                                                                                             | Term                          | References                                        |
|------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------|---------------------------------------------------|
| Research goal          | In this stage the basis for the research is layed down. The approach is to be determined and the scope of the research is formulated, making use of the strengths of qualitative methods to add value. | Scope creep                   | Roy and Searle, 2020                              |
|                        |                                                                                                                                                                                                        | Holistic approach             | Van Steenberg, 1990; Verschuren, 2001             |
|                        |                                                                                                                                                                                                        | Sensitizing concepts          | Boeije, 2010                                      |
|                        |                                                                                                                                                                                                        | Constructivist GT             | Charmaz, 2017                                     |
|                        |                                                                                                                                                                                                        | Discourse analysis            | Hodges et al., 2008                               |
|                        |                                                                                                                                                                                                        | Narrative analysis            | Franzosi, 1998                                    |
|                        |                                                                                                                                                                                                        | Phenomenology                 | Wojnar and Swanson, 2007                          |
|                        |                                                                                                                                                                                                        | Content analysis              | Stone et al., 1966                                |
| Data collection        | During this stage, regardless of the approach, the data collection is designed, making sure the data is collected such that it leads to a comprehensive answer to the research question.               | Qualitative data              | Carr et al., 2021                                 |
|                        |                                                                                                                                                                                                        | Observations                  | Angrosino and Rosenberg, 1994; Mulhall, 2003      |
|                        |                                                                                                                                                                                                        | Researcher triangulation      | Archibald, 2016; Patton, 1999                     |
|                        |                                                                                                                                                                                                        | Data & method triangulation   | Korstjens and Moser, 2018                         |
|                        |                                                                                                                                                                                                        | Interviews                    | Dearnly, 2005; Evers, 2015                        |
|                        |                                                                                                                                                                                                        | Topic list                    | Kallio et al., 2016; McGrath et al., 2019         |
|                        |                                                                                                                                                                                                        | Verbatim transcription        | McMullin, 2023                                    |
| Reflecting & Adjusting | It is important to critically assess the rigor of the study. By applying multiple strategies to improve different aspects of rigor, the quality of the research can be held high.                      | Rigor                         | Guba, 1981; Korstjens and Moser, 2018             |
|                        |                                                                                                                                                                                                        | Reflexivity (epistemological) | Mauthner and Doucet, 2003                         |
|                        |                                                                                                                                                                                                        | Reflexivity (personal)        | Palaganas et al., 2017                            |
|                        |                                                                                                                                                                                                        | Saturation                    | Fusch and Ness, 2015                              |
|                        |                                                                                                                                                                                                        | Theoretical sampling          | Coyne, 1997                                       |
|                        |                                                                                                                                                                                                        | Negative cases                | Henry, 2015                                       |
| Data analysis          | As soon as the first data is collected the analysis will start. Essential in the approach of constructivist grounded theory is to keep an open mind for unexpected information.                        | Open coding                   | Strauss and Corbin, 1990                          |
|                        |                                                                                                                                                                                                        | A priori codes                | Stemler, 2000                                     |
|                        |                                                                                                                                                                                                        | Axial coding                  | Scott and Medaugh, 2017; Strauss and Corbin, 1990 |
|                        |                                                                                                                                                                                                        | Selective coding              | Strauss and Corbin, 1990                          |
|                        |                                                                                                                                                                                                        | Core category                 | Giles et al., 2016                                |
|                        |                                                                                                                                                                                                        | Constant comparison           | Boeije, 2010                                      |
| Intervention effect    | In the final research stage, the comparison between experimental groups is conducted to determine the intervention's effect, if any.                                                                   | Types of theoretical models   | Sandelowski and Barroso, 2003                     |
|                        |                                                                                                                                                                                                        | Key principles of theories    | Glaser and Strauss, 1967                          |

comprehensive answer possible. In the worked example, GLS adjusted their plans several times; e.g., in the second wave of data collection they included more extensive interviews to get a better understanding of the motives and strategies of the participants, and in different phases of the analysis they moved back and forth between ignoring or enforcing *empathy* as a sensitizing concept or through a priori codes, to search for the right balance between guiding and structuring the data analysis and letting the data speak for itself without giving too much direction.

To elaborate a bit on this last point, the approach to data analysis in general and the sensitizing concept of *empathy*, specifically, was highly flexible. While theory suggests that a priori codes should be formulated for sensitizing concepts, the researchers chose not to do so, as they were concerned that this could lead to over-fitting the data. At the start of axial coding, they explored the fit of the sensitizing concept with the data, a common practice in qualitative research. Ultimately, the researchers made the decision to discard the concept again at a certain point during axial coding. This shows once more that qualitative studies involve ongoing reflection, iteration, and adaptation: even in a constructivist grounded theory approach, known to be relatively structured compared to other qualitative approaches. The decision-making process should be driven by the research goals and commitment to rigor and transparency.

Two limitations of our study are important to mention. In the second wave of data collection in the empirical example, GLS managed to interview only a small number of participants in the intervention group. Even when invited for the interview near the intervention location, most people preferred the offered option of doing the interview at another time, often incidentally during a control week, or they preferred an online interview. This changed their status to control group participants since they did not encounter the intervention shortly before the interview. Despite observing this throughout the data collection phase, and stimulating the interviewers to motivate people to be interviewed at the intervention location during an intervention week, the final number remained low. It is a clear example of reflecting and adjusting throughout the study (by adapting instructions to interviewers), but in this example it did not lead to more interviews in the experimental condition. We believe that negative impact was manageable and that the data was still rich enough to draw conclusions with regards to the absence of an effect of the interventions. Either way, the worked example did enable a thorough testing of all methodological stages, including the final comparison between intervention and control group, leading to the methodological recommendations presented in this paper. Though a

more exhausting experimental group might have enriched the data analysis even further, we think the steps identified would not have changed with more data.

Basing our paper largely on the worked example, however, has another limitation and that is the somewhat narrow look on the different opportunities that qualitative research methods offer. We discussed several approaches in relative depth (e.g., interviews, observation, the CGT approach) because they were applied in the worked example. And we left out others (e.g., focus groups, archival data, case studies) that were not applied in the worked example. In order to be more exhaustive, in Table 4.3 several additional approaches and methods are listed, including relevant literature references.

Following our recommendations, supplemented by the information in Table 4.3, researchers will gain the necessary tools to effectively employ qualitative methodologies for evaluating experimental effects. Compared to our example, it will be more straightforward to apply this approach in more controlled experimental settings, like in a laboratory environment. An evident challenge highlighted in the worked example was the variance in participant engagement with the intervention, a common occurrence in field experiments where researchers contend with less control compared to lab experiments. Augmenting or even replacing traditional quantitative methods with qualitative approaches could lead to unforeseen revelations regarding intervention effects.

In conclusion, our paper provides guidelines for the qualitative evaluation of experimental designs. While these guidelines are not all-encompassing, they provide a useful starting point for researchers looking to incorporate qualitative methods in their experimental research. We hope to have convinced readers that sometimes experiments in social and behavioral research can be studied in a more meaningful and more insightful way if one either changes to qualitative methods, or adds a qualitative evaluation of results to the more traditional quantitative way of measuring intervention effects.

## Acknowledgments

We are grateful to Prof. Dr. Denise de Ridder for setting up the bigger context study and allowing for the qualitative study that served as a worked example in this study to be executed. We thank Dr. Marie-Louise Glebbeek for offering a critical reflection during the design and execution of the empirical study. We thank Efe Adu, Timo Keijzer, Jamie Keurntjes, Jonas van Oosten, & Danja Roelofs for their contributions to the data collection. We used ChatGPT for textual improvements.

The preregistration for this study can be accessed via this link:  
<https://archive.org/details/osf-registrations-knb2d-v1>

## Bibliography

- Angrosino, M., & Rosenberg, J. (1994). Observations on observation. In *The sage handbook of qualitative research* (4th, pp. 467–478). SAGE.
- Archibald, M. (2016). Investigator triangulation: A collaborative strategy with potential for mixed methods research. *Journal of Mixed Methods Research*, 10, 228–250. <https://doi.org/10.1177/1558689815570092>
- ATLAS.ti Scientific Software Development GmbH. (2022). Atlas.ti 22 windows [Available from <https://atlasti.comm>].
- Boeije, H. (2010). *Analysis in qualitative research*. SAGE publications Ltd. <https://doi.org/10.5785/26-2-24>
- Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research*, 6, 97–113. <https://doi.org/10.1177/1468794106058877>
- Carr, D., Boyle, E., Cornwell, B., Correll, S., Crosnoe, R., J., F., & Waters, M. (2021). *The art and science of social research*. W.W. Norton.
- Charmaz, K. (2017). Constructivist grounded theory. *The Journal of Positive Psychology*, 12, 299–300. <https://doi.org/10.1080/17439760.2016.1262612>
- Chauvette, A., Schick-Makaroff, K., & Molzahn, A. (2019). Open data in qualitative research. *International Journal of Qualitative Methods*, 18, 1–6. <https://doi.org/10.1177/1609406918823863>
- Class, B., De Bruyne, M., Willemin, C., Donzé, D., & Claivaz, J.-B. (2021). Towards open science for the qualitative researcher: From a positivist to an open interpretation. *International Journal of Qualitative Methods*, 20, 1–15. <https://doi.org/10.1177/16094069211034641>
- Cohen, B., & Lea, R. (2004). *Essentials of statistics for the social and behavioral sciences*. John Wiley & Soncs, Inc.
- Corbin, J., & Strauss, A. (1990). Grounded theory research: Procedures, canons and evaluative criteria. *Qualitative Sociology*, 13, 3–21. <https://doi.org/10.1007/BF00988593>

- Coyne, I. (1997). Sampling in qualitative research. purposeful and theoretical sampling; merging or clear boundaries? *Journal of Advanced Nursing*, 26, 623–630. <https://doi.org/10.1046/j.1365-2648.1997.t01-25-00999.x>
- Cuff, B., Brown, S., Taylor, L., & Howat, D. (2016). Empathy: A review of the concept. *Emotion Review*, 8(2), 144–153. <https://doi.org/10.1177/1754073914558466>
- De Ridder, D., Aarts, H., Benjamins, J., Glebbeek, M., Leplaa, H., Leseman, P., Potgieter, R., Tummers, L., & Zondervan-Zwijnenburg, M. (2022). “keep your distance for me”: A field experiment on empathy prompts to promote distancing during the covid-19 pandemic. *Journal of Community & Applied Social Psychology*, 32, 755–766. <https://doi.org/10.1002/casp.2593>
- Dearnly, C. (2005). A reflection on the use of semi-structured interviews. *Nurse Researcher*, 13, 19–28. <https://doi.org/10.7748/nr2005.07.13.1.19.c5997>
- Denzin, N., & Lincoln, Y. (2017). *The sage handbook of qualitative research*. Sage Publications, Inc.
- Evers, J. (2015). *Kwalitatief interviewen: Kunst en kunde*. Boom/Lemma.
- Franzosi, R. (1998). Narrative analysis, or why (and how) sociologists should be interested in narrative. *Annual Review of Sociology*, 24, 517–554. <https://doi.org/10.1146/annurev.soc.24.1.517>
- Fusch, P., & Ness, L. (2015). Are we there yet? data saturation in qualitative research [Retrieved from <https://www.proquest.com/scholarly-journals/are-we-there-yet-data-saturation-qualitative/docview/1721368991/se-2>]. *The Qualitative Report*, 20, 1408–1416.
- Giles, T., De Lecay, S., & Muir-Cochrane, E. (2016). Coding, constant comparisons, and core categories: A worked example for novice constructivist grounded theorists. *Advances in Nursing Science*, 39, 29–44. <https://doi.org/10.1097/ANS.0000000000000109>
- Glaser, B. (1978). *Theoretical sensitivity*. Mill Valley, CA: The Sociology Press.
- Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine.
- Glebbeek, M., Leplaa, H. J., & Soppe, K. F. B. (2024). The power of prompting: A qualitative experiment on prompting empathy [[Manuscript submitted for publication]. Department of Methods and Statistics, Utrecht University.]

- Gribbons, B., & Herman, J. (1996). True and quasi-experimental designs. *Practical Assessment, Research, and Evaluation*, 5(14). <https://doi.org/10.7275/fs4z-nb61>
- Guba, E. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Resources Information Center Annual Review Paper*, 29, 75–91. <https://doi.org/10.1007/BF02766777>
- Haven, T., & Van Grootel, L. (2019). Preregistering qualitative research. *Accountability in Research*, 26, 229–244. <https://doi.org/10.1080/08989621.2019.15801>
- Henry, P. (2015). Rigor in qualitative research: Promoting quality in social science research [Retrieved on 29 March 2023 from <http://www.isca.me/rjrs/archive/v4/iIVC-2015/6.ISCA-IVC-2015-20SHS-11.php>]. *Research Journal of Recent Sciences*, 4, 25–28.
- Henwood, K., & Pidgeon, N. (1995). Remaking the link: Qualitative research and feminist standpoint theory. *Feminism & Psychology*, 5, 7–30. <https://doi.org/10.1177/0959353595051003>
- Hodges, B., Kuper, A., & Reeves, S. (2008). Qualitative research: Discourse analysis. *British Medical Journal*, 337, 570–572. <https://doi.org/10.1136/bmj.a879>
- Humphreys, M., Sanchez de la Sierra, R., & Van der Windt, P. (2013). “fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration.” *Political Analysis*, 21, 1–20. <https://doi.org/10.1093/pan/mps021>
- Kallio, H., Pietilä, A.-M., Johnson, M., & Kangsniemi, M. (2016). Systematic methodological review: Developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, 72, 2954–2965. <https://doi.org/10.1111/jan.13031>
- Kearney, M. (2001). Levels and applications of qualitative research evidence. *Research in Nursing & Health*, 24, 145–153. <https://doi.org/10.1002/nur.1017>
- Korstjens, I., & Moser, A. (2018). Series: Practical guidance to qualitative research. part 4: Trustworthiness and publishing. *European Journal of General Practice*, 24, 120–124. <https://doi.org/10.1080/13814788.2017.1375092>
- Kratochwill, T., Hitchcock, J., Horner, R., Levin, J., Odom, S., Rindskopf, D., & Shadish, W. (2010). Single-case design technical documentation [Retrieved at April 20 2023 from <https://eric.ed.gov/?id=eD510743>].

- Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. Sage. [https://doi.org/10.1016/0147-1767\(85\)90062-8](https://doi.org/10.1016/0147-1767(85)90062-8)
- List, J. (2007). Field experiments: A bridge between lab and naturally occurring data. *Advances in Economics, Analysis, BAH Policy*, 6, 1–45. <https://doi.org/10.2202/1538-0637.1747>
- Luo, Y., Sun, F., Jiang, L., & Zhang, A. (2019). The stress and coping experiences among chinese rural older adults in welfare housing: Through the lens of life review. *Aging and Mental Health*, 29, 1086–1094. <https://doi.org/10.1080/13607863.2019.1612322>
- Mauthner, N., & Doucet, A. (2003). Reflexive accounts and accounts of reflexivity in qualitative data analysis. *Sociology*, 37, 413–431. <https://doi.org/10.1177/00380385030373002>
- Maxwell, J. (2009). *The sage handbook of applied social research methods*. SAGE Publications, Inc. <https://doi.org/10.4135/9781483348858.n7>
- McGrath, C., Palmgren, P., & Liljedahl, M. (2019). Twelve tips for conducting qualitative research interviews. *Medical Teacher*, 41, 1002–1006. <https://doi.org/10.1080/0142159X.2018.1497149>
- McMullin, C. (2023). Transcription and qualitative methods: Implications for third sector research. *Voluntas*, 34, 140–153. <https://doi.org/10.1007/s11266-021-00400-3>
- Mulhall, A. (2003). In the field: Notes on observation in qualitative research. *Journal of Advanced Nursing*, 41, 306–313. <https://doi.org/10.1046/j.1365-2648.2003.02514.x>
- Naber, A. (2015). Qualitative experiment as a participating method in innovation research. *Historical Social Research*, 40, 233–257. <https://doi.org/http://www.jstor.org/stable/24583154>
- Palaganas, E., Sanchez, M., Molintas, V., & Caricativo, R. (2017). Reflexivity in qualitative research: A journey of learning. *Qualitative Report*, 22, 426–438. <https://doi.org/10.46743/2160-3715/2017.2552>
- Palsola, M., Renko, E., Kostamo, K., Lorencatto, F., & Hankonen, N. (2020). Thematic analysis of acceptability and fidelity of engagement for behaviour change interventions: The let’s move it intervention interview study. *British Journal of Health Psychology*, 25, 772–789. <https://doi.org/10.1111/bjhp.12433>
- Paluck, E. (2010). The promising integration of qualitative methods and field experiments. *The Annals of the American Academy*, 628, 59–71. <https://doi.org/10.1177/0002716209351510>

- Patton, M. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research, 34*, 1189–1208. <https://doi.org/PMC1089059>
- Pelak, C., & Duncan, S. (2016). Using a social science-fictional play to teach about global capitalism and macro-structural systems in introduction to sociology. *Teaching Sociology, 45*, 1–13. <https://doi.org/10.1177/0092055X16663804>
- Prowse, M., & Camfield, L. (2013). Improving the quality of development assistance: What role for qualitative methods in randomized experiments? *Progress in Development Studies, 13*, 51–61. <https://doi.org/10.1177/146499341201300104>
- QSR International Pty Ltd. (2020). Nvivo [Available from <https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home>].
- Reiss, A. (1971). Systematic observation of natural social phenomena. *Sociological Methodology, 3*, 3–33. <https://doi.org/10.2307/270816>
- Ritchie, J., Lewis, J., Mc Nuaghton Nicholls, C., & Ormston, R. (2013). *Qualitative research practice: A guide for social science students and researchers*. Sage Publications, Inc.
- Roe, B., & Just, D. (2009). Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics, 91*(5), 1266–1271. <https://doi.org/10.1111/J.1467-8276.2009.01295.X>
- Rosman, G., & Wilson, B. (1996). Number and words: Combining quantitative and qualitative methods in a single large-scale study. *Evaluation Review, 9*(5), 627–643. <https://doi.org/10.1177/0193841X8500900505>
- Roy, S., & Searle, M. (2020). Scope creep and purposeful pivots in developmental evaluation. *Canadian Journal of Program Evaluation, 35*, 92–103. <https://doi.org/10.3138/CJPE.56898>
- Sandelowski, N., & Barroso, J. (2003). Classifying the findings qualitative studies. *Qualitative Health Research, 13*, 905–923. <https://doi.org/10.1177/1049732303253488>
- Schmuckler, M. (2001). What is ecological validity? a dimensional analysis. *Infancy, 2*, 419–436. <https://doi.org/10.1207/S15327078IN0204\02>
- Scott, C., & Medaugh, M. (2017). Axial coding. *The International Encyclopedia of Communication Research Methods, 1*–2. <https://doi.org/10.1002/9781118901731.iecrm0012>

- Smaling, A. (1994). The pragmatic dimension: Paradigmatic and pragmatic aspects of choosing a qualitative or quantitative method. *Quality & Quantity*, 28, 233–249. <https://doi.org/10.1007/BF01098942>
- Stangor, C. (2014). *Research methods for the behavioral sciences*. Cengage Learning.
- Stemler, S. (2000). An overview of content analysis. *Practical Assessment, Research, and Evaluation*, 7, 137–146. <https://doi.org/10.7275/z6fm-2e34>
- Stone, P., Dunphy, D., Smith, M., & Ogilvie, D. (1966). *The general inquirer: A computer approach to content analysis*. MIT Press. <https://doi.org/10.2307/1161774>
- Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Sage.
- Tijmstra, J., & Boeije, H. (2011). *Wetenschapsfilosofie in de context van de sociale wetenschappen*. Boom.
- Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociological Theory*, 30, 167–186. <https://doi.org/10.1177/0735275112457914>
- Van Grootel, L., Balachandran Nair, L., Klugkist, I., & Van Wesel, F. (2020). Quantitizing findings from qualitative studies for integration in mixed methods reviewing. *Research Synthesis Methods*, 15(11), 413–425. <https://doi.org/10.1002/jrsm.1403>
- Van Steenberg, B. (1990). Potential influence of the holistic paradigm on the social sciences. *Futures*, 10, 1071–1083. [https://doi.org/10.1016/0016-3287\(90\)90008-6](https://doi.org/10.1016/0016-3287(90)90008-6)
- VERBI Software. (2021). Maxqda 2022 [Available from maxqda.com].
- Verschuren, P. (2001). Holism versus reductionism in modern social science research. *Quality & Quantity*, 35, 389–405. <https://doi.org/10.1023/A:1012242620544>
- Vors, O., Marqueste, T., & Mascret, N. (2018). The trier social stress test and the trier social test for groups: Qualitative investigations. *PLoS ONE*, 13. <https://doi.org/10.1371/journal.pone.0195722>
- Williams, M., & Moser, T. (2019). The art of coding and thematic exploration in qualitative research [Retrieved April 14, 2023, from <http://www.imrjournal.org/uploads/1/4/2/8/14286482/imrv15n1art4.pdf>]. *International Management Review*, 15, 45–55.
- Wojnar, D., & Swanson, K. (2007). Phenomenology. *Journal of Holistic Nursing*, 25, 172–180. <https://doi.org/10.1177/0898010106295172>

## CHAPTER 5

### Prompting in practice: A qualitative field experiment on prompting empathy

Glebbeek, M.-J., Leplaa, H.J., & Soppe, K.F.B. (2025). Prompting in Practice: A Qualitative Field Experiment on Prompting Empathy. Under review.

#### Abstract

The relationship between prompts and behavioral outcomes has been studied in clinical trials and quantitative research, but qualitative methods may offer deeper insight into underlying mechanisms. A qualitative experiment during the COVID-19 pandemic at Dutch higher education institutions examined how empathy-inducing prompts influenced behavior. Using an A-B design, the study alternated between control (A) and experimental (B) weeks, employing prompts to induce empathy during experimental weeks. Data came from predominantly white, highly educated students, collected through systematic and unsystematic observations complemented by short (Wave 1) and long (Wave 2) interviews. Findings revealed the complexity of student behavior, showing that while attitudes toward COVID-19 restrictions shaped intended behavior, these intentions did not consistently translate into actions. Contextual factors further influenced this relationship. Moreover, minimal differences between control and experimental groups suggested that the prompts had little measurable impact on behavioral outcomes.

*Keywords:* Prompting; Empathy; Behavioral outcomes; Qualitative methods; Experiment; COVID-19

---

Author contribution: KS, MG, & HL developed the theoretical framework for the study. HL & MG designed the study. MG & HL supervised data collection, KS collected data. HL conducted the entire data analysis, MG was involved in the initial stages of data analysis, KS in the later stages of data analysis. MG, HL, & KS wrote and revised the article.

## 5.1 Introduction

Empathy is a multifaceted concept that has been defined in many ways (Cuff et al., 2016). There seems to be a consensus that empathy itself does not have a behavioral component. Using the definition of Cohen and Strayer (1996, p.988) we define empathy as “The ability to understand and share in another’s emotional state or context.” Feelings of empathy make people motivated to act in favor of those they feel empathy towards. Favero and Pedersen (1996, p.2) define prosocial motivation as “the desire to protect or improve the well-being of other people.” They state that research shows that prosocial motivation drives decisions that are beneficial to the collective and for attaining collective goals (Favero & Pedersen, 1996). Orientations towards increasing the welfare of others, especially the vulnerable others, are the result of altruistic motivation evoked by empathy (Favero & Pedersen, 1996). Precisely the emphatic response is the mechanism that explains how feelings of empathy translate into prosocial behavior (Oliveira-Silva & Gonçalves, 2011).

Multiple studies have shown that experiencing empathy affects motivation to act in certain ways (e.g. Martinez and Knee, 2024; Pfattheicher et al., 2020) and has the potential to result in change in actual behavior (e.g. Grant & Hoffmann 2011; De Ridder et al., 2022). Moreover, empathy levels appear to change during the life course. Schieman and Van Gundy, 2000 found that older adults reported lower empathy levels than adolescents. Gutiérrez-Cobo et al., 2023 also highlighted that higher educational levels can buffer the negative effects of aging on empathy, suggesting that the cognitive processes related to feeling empathy diminishes less in well-educated older adults. Moreover, Malin et al., 2014, based on a sample of teens and adolescents, emphasized that during adolescence, empathy development is crucial for role exploration and purpose, indicating that empathy evolves with age and is influenced by social contexts. These findings collectively suggest an inversed U-shape on the relation between age and level of empathy, with the highest impact on social behavior during adulthood and lower levels in early and later life stages. Studies on emphatic responses are mostly based on clinical trials (such as Marci et al. 2007) and the few exceptions are also quantitative in nature (e.g. Pfattheicher et al. 2020), with a big focus on whether behavioral change occurs. The quantitative approach leaves the question of how behavioral change comes about, especially in real life settings, largely untouched.

Nowadays, there is a large body of research showing that people’s behavior can be altered through subtle interventions called prompts (see for relevant

systematic reviews e.g. Alkhaldi et al. 2016; De Leon et al. 2014). Prompts are “verbal or written antecedent messages that designate desirable target behaviors” (Lehman and Geller, 2004, p.19). Research on prompts is widespread, spanning many disciplines such as health science (e.g., Maltinsky & Swanson, 2020; Prinsen et al., 2013), environmental research (Brown et al., 2010; Wichmann et al., 2010), and public policy (e.g., Whitehead et al., 2014). Most studies researching prompting can be found in the field of psychology (e.g. Leoniak and Cwalina 2019; De Ridder et al. 2022; Pfattheicher et al. 2020).

During the COVID-19 pandemic, research has been carried out about empathy-inducing cues and its influence on prosocial behavior, such as adhering to social distancing or wearing face masks (e.g. Favero & Pederson 1996; Martinez and Knee 2024; Pfattheicher et al. 2020; Lunn et al. 2019). Outcomes of these studies, concerning the effectiveness of empathy-inducing cues on prosocial behavior and the adherence to COVID-19 restrictions, are mixed. Favero and Pedersen, 1996 stated that prosocial and empathy appeals were ineffective in social distancing compliance. Moreover, De Ridder et al., 2022 reported inconclusive results in a study that used empathy-prompts to foster social distancing.

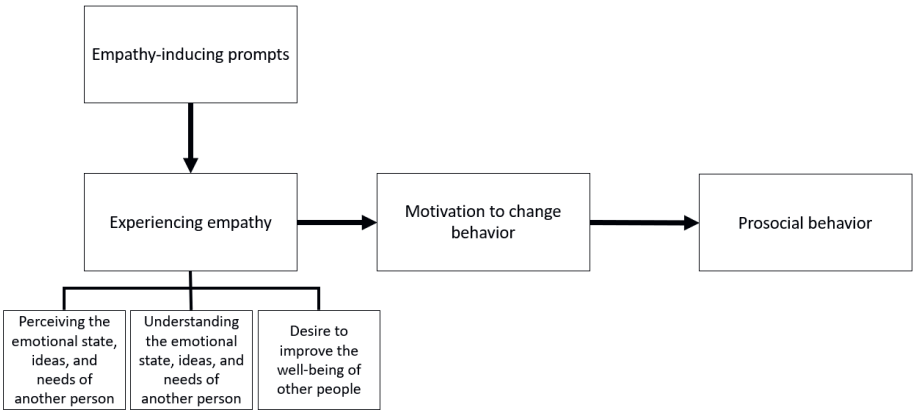
### 5.1.1 The present study

This study aims for a more insightful understanding of the processes that take place between receiving a prompt and performed behavior. Despite the large body of research from clinical trials and other quantitative studies showing the effectiveness of prompting, in real-world settings there is limited research demonstrating that empathy prompts may lead to change in actual behavior. To our knowledge, no qualitative studies exist studying the underlying mechanisms at play. Hence, an experiment with qualitative data was conducted during the COVID-19 pandemic, specifically targeting students, at Utrecht Science Park (USP), the campus of Utrecht. This allows us to get a deeper understanding of the processes at play for this homogenous group of young, highly educated people. In sum, this study investigates two key questions, first: What are the mechanisms underlying the emergence of (prosocial) behavior in the context of COVID-19 restrictions, and second: How do empathy-inducing prompts influence behavioral intentions and outcomes among Dutch higher education students?

Figure 5.1 visualizes the hypothesized mechanisms leading to (prosocial) behavior. In this model, prompts are expected to elicit empathy in participants—defined as the capacity to perceive and understand another person’s

emotional state, ideas, and needs, along with a desire to enhance their well-being. This empathetic response is hypothesized to motivate individuals to engage in more prosocial behavior.

**Figure 5.1: The theoretical model visualizing how prompts could lead to prosocial behavior by inducing empathy.**



## 5.2 Method

This study is part of a larger research project on empathy-inducing prompts and behavioral outcomes, conducted during the COVID-19 pandemic at Utrecht Science Park. The larger project analyzed quantitative data (camera recordings) to assess distancing behavior (De Ridder et al. 2022). Our study adds a qualitative perspective through observations and interviews, providing insights into behavioral patterns and motivations behind compliance with COVID-19 restrictions. The multidisciplinary core team (psychologist, sociologist, anthropologist) ensured a holistic approach.

### 5.2.1 Design

A field experiment was conducted in two waves: late fall 2020 (Wave 1) and late spring 2021 (Wave 2), during partial lockdown. A pilot week preceded Wave 1 data collection to test procedures. An A-B design was used, alternating between control weeks (A) and intervention weeks (B). The intervention in B-weeks consisted of empathy-inducing prompts with three components:

1. A social robot greeting people and reminding them to keep distance (Figure 5.2);
2. Posters of students/staff with empathy-based messages (e.g., “I have asthma. Keep your distance for me,” 5.3);
3. Movie clips with audio messages from the same students/staff.

The prompts were placed at three key locations on campus: outside a college hall, at the entrance of a lecture hall, and at a large lecture room entrance. Wave 1 included three A-B sequences (six weeks), and Wave 2 included four A-B sequences (eight weeks), all at the same campus locations.

**Figure 5.2: The social robot greeting people and encouraging them to keep distance.**

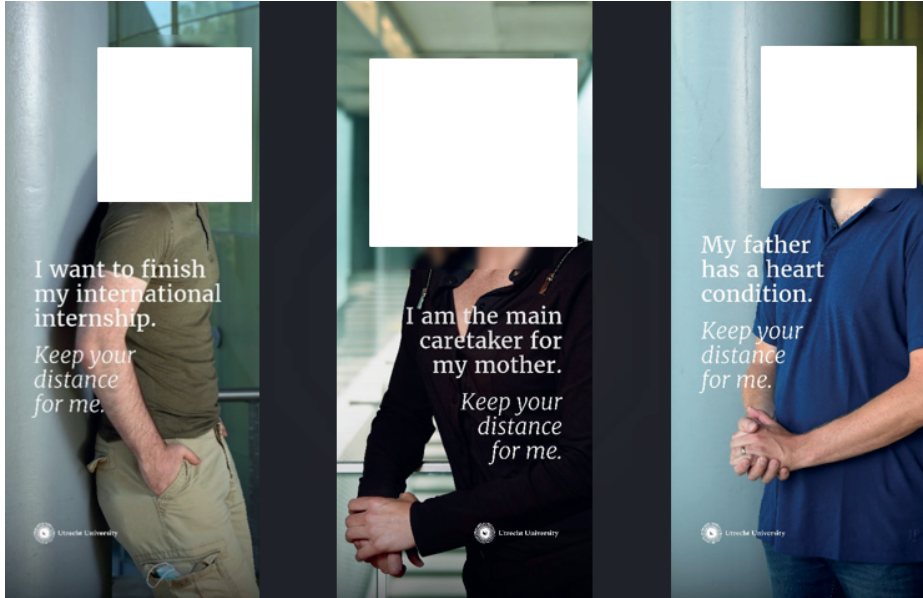


### 5.2.2 Participants and Procedure

Data collection in Wave 1 combined systematic and non-systematic observations to assess whether individuals adjusted their behavior in response to empathy-inducing prompts and to identify contextual factors influencing adherence to COVID-19 restrictions. Six researchers used a checklist to record whether people noticed the prompts, maintained distance, and walked alone or in groups. The checklist also allowed for additional non-systematic information in the form of fieldnotes, which provided detailed descriptions of behavior, obstacles to distancing, and situational changes. Observations totaled 52 hours in the control group and 70 hours in the intervention group. Researcher triangulation (six researchers) minimized individual bias (Archibald, 2016; Patton, 1999).

To complement observations, we conducted short interviews in the final week

**Figure 5.3:** Three of the posters of student and staff models, white boxes added by us for privacy reasons, with a text expressing a prompt for empathy-based distancing.

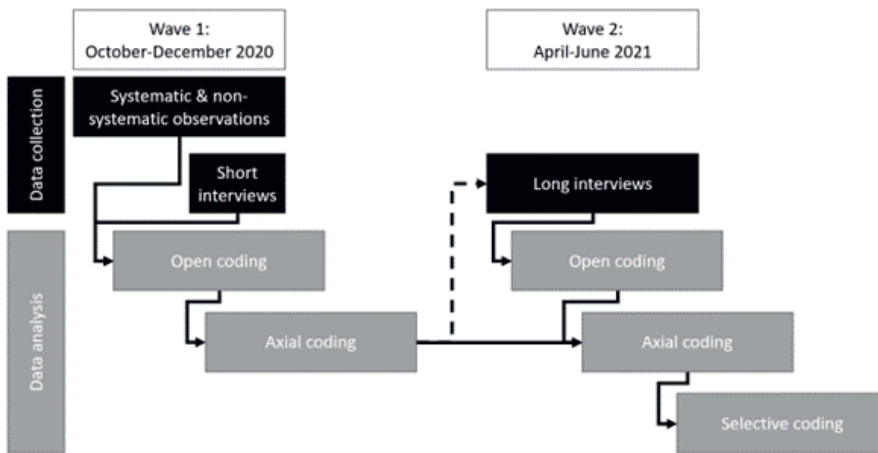


of Wave 1. Six researchers carried out 5-minute walk-along interviews with students they observed, guided by a predefined topic list (Appendix 5.A). The total of 42 interviews was equally split between a site close to the intervention and a square in front of some buildings quite far away from the intervention. The latter served as control location for the short interviews. These short interviews were subsequently transcribed intelligent verbatim to preserve meaning (McMullin, 2023). The combination of observations and short interviews enabled triangulation, providing a holistic view of the intervention's effects.

However, analysis of Wave 1 interviews revealed insufficient depth to fully understand motivations. Therefore, Wave 2 focused on long, semi-structured interviews of about 45 minutes. Conducted over eight weeks, these long interviews addressed themes derived from preliminary analysis of Wave 1 data, such as social life, home situation, and study-related contact (Appendix 5.B). In total, 39 long interviews were conducted by five researchers across three settings: near interventions (experimental group,  $N = 6$ ), another campus building (control,  $N = 13$ ), online (control,  $N = 3$ ), and the intervention site during control weeks (control,  $N = 17$ ). These interviews were also transcribed intelligent verbatim.

By combining observations, short interviews, and later long, in-depth, interviews, the study aimed to capture both behavioral outcomes and the underlying attitudes, intentions, and motivations related to COVID-19 restrictions. The sample largely consisted of young, highly educated students, making it relatively homogeneous. Due to the set-up of data collection, students may have been observed multiple times but were interviewed only once. Figure 5.4 illustrates the study's workflow, including the different stages of data collection and analysis. Materials are available on OSF (<https://osf.io/e4zfw/>).

**Figure 5.4: Workflow of the study.**



### 5.2.3 Ethics statement

The Ethics Committee of the Faculty of Social and Behavioral Sciences, Utrecht University approved the study (file 20-479). Information signs about data collection were posted at data collection sites and linked to the university website. Before interviews, participants gave informed consent. Short interviews and observations were anonymous, and transcripts of long interviews contained no identifying details. No deception was used.

### 5.2.4 Transparency and Openness

This study was not preregistered. We report how sample size was determined, any data exclusions, all manipulations, and all measures. Data collection and analysis followed a reflexive approach, with early analysis guiding theoretical

sampling. Although loss of context limits interpretation in case of data re-use, all data and materials (checklists, topic lists) are publicly available on OSF. Furthermore, the axial coding scheme and created queries can be found on OSF.

### 5.2.5 Analyses

We used the Grounded Theory approach (Corbin & Strauss, 1990; Glaser, 1978) to analyze the data in NVivo 2020. Differentiation between conditions was introduced only after the initial steps of analysis. We assumed the core process leading to behavior was shared across participants, though certain aspects could differ. These differences were examined by assigning labels to cases and running queries after building the empirical model.

The analysis began with open coding by six researchers, followed by axial coding conducted by two of them with two additional researchers. Initially, only Wave 1 data were available and analyzed up to selective coding. When this proved insufficient for the study's objectives, additional interview data were collected in Wave 2, and the final analysis was based on both waves.

During axial coding, we distinguished between 'motivations' and 'strategies', found in both observational and interview data. Motivations referred to reasons behind behaviors, while strategies described the behaviors themselves. Within strategies, we separated 'intended behavior' from 'actual behavior'. Initially, 'empathy' was used as a sensitizing concept in this stage as one of the motivations. It was quickly dropped, since it led to overfitting of the data. The coding tree was restructured into (sub)themes forming the empirical model. Table 5.1 illustrates this process, showing axial codes alongside related sub-codes, although some were refined later. Table 1 is an illustration, and not an extensive list of all codes and fragments. The full axial coding scheme can be found on OSF.

Selective coding and the identification of differences between intervention and control groups were carried out by three researchers. This process examined data from multiple perspectives. Only comparisons that were deemed relevant are presented in the results, full output from the queries can be found on OSF. The second wave was crucial for constructing the final empirical model, though data from both waves were rigorously analyzed at each stage. Based on the set-up of this study, we would expect that differences between data from intervention and control groups would be clear. Thus, identifying the group that any interview

Table 5.1: A selection of subthemes, open codes, and fragments to serve as illustrations for the coding scheme resulting from the axial coding phase.

| Theme             | Subtheme                | Open code                                                 | Fragment                                                                                                                                                                                                                                                                                                                                   |
|-------------------|-------------------------|-----------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Considerations    | Perceived risk for self | No worries for own health                                 | ‘For example, we all still have parents that we want to protect or grandparents or just people in our environment so, usually we are, like it doesn’t matter that much to ourselves. We are young, things will probably turn out fine, but also regarding the people we love, we really want to stick to it as much as possible.’ DL240603 |
|                   |                         | COVID-fatigue                                             | ‘But I also kind of need some human interaction and I want to hug, at least with two people in the Netherlands. And that that was a small risk I was willing to take’ DL240601                                                                                                                                                             |
| Intended behavior | Agreements with others  | Adherence to rules becomes diluted                        | ‘I must say that in the beginning I was very strict about that. After a year and a half, that fades a bit.’ JL080601                                                                                                                                                                                                                       |
|                   |                         | Making agreements with parents about adhering to measures | ‘We have informally discussed what we think of the measures and whether others or many people still see them. (..) “How do you feel about it? Also, are you vulnerable or are you very afraid of it or not?”                                                                                                                               |
|                   |                         |                                                           | EL190503                                                                                                                                                                                                                                                                                                                                   |
| Context factors   | Interchangeability      | Breaking rules only in small groups                       | ‘If I break the rules, I don’t do it with 15 different people, I just do it with the two or three’ MB090601                                                                                                                                                                                                                                |
|                   | Noticing prompts        |                                                           | ‘(..) Now she sees the robot and look at it together with 2 fellow students. They wave at the robot, hoping it would react. When it responds they are disappointed it talks about COVID-19 regulations’ JmB12110413                                                                                                                        |

|                                    |                      |                                    |                                                 |                                                                                                                                                                                                                                                                                                                                                                                                                               |
|------------------------------------|----------------------|------------------------------------|-------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Behavior<br>others<br>dis-<br>play | Actual be-<br>havior | Mimicking<br>behavior of<br>others | Behavior influenced<br>by family                | 'Female, walks with a female friend, at first at 1.5m distance. Takes some trouble to walk around their friend when they stop in their tracks. As the friend starts walking again they walk close to the observe. Observee doesn't try to reassure the 1.5m distance now. (..)' MaA16100105                                                                                                                                   |
|                                    |                      |                                    |                                                 | 'Sometimes my mother, for example, wants to give me a hug when I return home to Utrecht. And then I sometimes say: Oh, do you mind the distance. And then she says: yes, but I've already had a shot, I can do it now. And then, uh, yes, I let it go I guess.' JL040601                                                                                                                                                      |
|                                    |                      |                                    | Behavior influenced<br>by friends               | 'I think you just have to look at that very much, whatever a person thinks about it. For example, I have a friend who really likes to keep his distance. And I respect that. I make sure that when we uhm, when he comes to me, that was the other week. Then I also make sure that we so nicely that he sits on the couch. That I grab an extra chair. And that in that way we do really say take it into account.' EL020602 |
|                                    |                      | Noticing<br>prompts                | Glances at inter-<br>vention while pass-<br>ing | 'Female, Has a mouth mask in her mouth. Walks while eating. Looks at the intervention as she walks.' JmB12110208                                                                                                                                                                                                                                                                                                              |
|                                    |                      |                                    | Discuss interven-<br>tion                       | 'Female, approaches from the cafeteria with 3 friends, they all keep 1.5 meters away. She notices the robot, it's not on, but she says something about it, makes the comment whether they will walk in again hoping it turns on, don't do that.' MbB19100402                                                                                                                                                                  |

or observation belonged to should be possible while knowing the results of this study, but only if the interventions are effective.

To assess the prevalence of (sub)themes across conditions, we used NVivo classifications linking data to locations, weeks, and researchers. Initial queries were guided by the theory that empathy prompts influence behavior, focusing on motivations. Narratives were developed for students in both groups to highlight differences. This approach provided insight into adherence and non-adherence to COVID-19 restrictions, distinguishing between control and experimental data while clarifying behavioral patterns.

## 5.3 Results

The qualitative analysis of the data, obtained through observations and interviews, showed that, the process behind (intended) behavior with regards to COVID-19 restrictions and how this process may be affected by the prompts, leads to the identification of four main aspects: 1) considerations and beliefs, which result in 2) intended behavior, which in turn will, influenced by 3) context factors, lead to 4) actual behavior. These aspects form the basis of the empirical model presented in Figure 5.5, provides the structure for describing the results in the subsections below. Throughout the section, *italic words* refer to important subthemes.

### 5.3.1 Considerations and beliefs

When students were confronted with the COVID-19 pandemic and the subsequent COVID-19 restrictions (e.g., social distancing, hygiene precautions, etc.) everyone had to consider how to deal with and adapt to this situation. In the data a few reoccurring considerations or beliefs that influenced students' intended behavior were identified. First, students weighed the possibility of getting infected themselves. As our sample consists of students, mostly in the age between 18 and 25 years old, who were at least healthy enough to visit the campus, they were not the group most at risk. During the interviews, students showed awareness of this and felt that it was not essential to adhere to COVID-19 restrictions for their personal health, see for instance the quote after the next paragraph of DR1006O1. Some students admitted the risk but considered it to be minor. Others downplayed the risk for themselves.

Second, most students in our study did acknowledge the increased severity of health risks for others and expressed a genuine willingness to modify their

behavior accordingly. One student expressed her consideration of conforming to the COVID-19 restrictions when meeting others:

*‘How terrible would it be if you caused someone else to die or become extremely ill? My mother had it at the beginning of the year and had to stay home for three months. So, I really wouldn’t want, say, that I do this to anyone else. So, that’s the main thing of course. See, you think. It doesn’t make me that sick, I’m very young. I don’t think it matters much, so it’s really for other people.’* (DR1006O1)

The older generations, specifically their parents and grandparents, were the primary focus of students’ concerns. They expressed a desire to adhere to the COVID-19 restrictions out of concern for their (grand)parents because they perceived them as more vulnerable. Additionally, some students have (grand)parents who are experiencing health issues. One of the students explained: *‘We all have parents or grandparents who we want to protect, (..) when it comes to the people who are precious to us, we want to adhere to the restrictions as much as possible’* (DR2406L3). This initial manifestation underscores the presence of empathy, as students demonstrate the ability to adjust their behavior even when it is not directly beneficial to them.

Third, another common rationale used to determine (intended) behavior was a presumed immunity to COVID-19. Students expressed that they deemed their behavior safe, since the people involved had been vaccinated or had recovered from a recent COVID-19 infection. And, since they were now immune for at least a short period of time, they felt it was safe to disregard some of the COVID-19 restrictions: *‘Take my parents for instance... They’re both vaccinated now. So now I’ve become a bit more lenient.’* (DL2406O3).

Last, students expressed that their attitudes regarding the COVID-19 restrictions changed over time. Some have referred to this as ‘COVID-fatigue’, alluding to a growing reluctance to adhere to COVID-19 restrictions as time passed. While students were initially willing to strictly adhere to all rules and regulations, they gradually developed a desire to get physically close to others and noticed a declining degree of patience and understanding for the COVID-19 restrictions. One student explained how this process worked in their shared living situation:

*‘Well, I know how it started. At first, we had strict rules for a while, like no visitors at all. That was for one of my housemates, they moved out now, but*

*they cared for their parent. But at some point, people needed to have contact with others, cause some had a girl- or boyfriend. And then we chose that [more lenient] way that you could have a visitor' (JT2005L2).*

Moreover, over time, students became less fearful of (getting infected with) COVID-19, partially because they realized that they were not the group at risk. At the same time, many students indicated that they felt the need to have contact with others, one student illustrated this with a contemplation about adhering to the rules and regulations: 'It was really a bit of a consideration of okay you know that you have a greater chance of contracting corona by meeting with a group of six people. But it wasn't for my mental health, yes, I think it had deteriorated so much that it wouldn't have been very good. So, then it is necessary' (DR2406L2).

In conclusion, we discovered that while many students saw little risk for themselves, they were concerned about others, such as elderly people, their parents, or other vulnerable people in their surroundings. Over time, as more people recovered from COVID-19 infections or received vaccinations, they were more reluctant to adhere to the restrictions, especially due to a need for more social interaction to foster their mental health.

### 5.3.2 Intended behavior

The data revealed that individuals had specific intentions regarding their behavior. Students expressed personal convictions or agreements about adherence to COVID-19 restrictions. While some students disclosed explicit agreements with their families or housemates, others mentioned more implicit or absent agreements. These intentions were formulated prior to social interactions, representing a chronological pattern where ideas and agreements were established prior to the actual encounters.

As many students lived in student accommodations with other students, most agreements students talked about were made in these settings. A student described one of the agreements they made in their student housing: '*I live with four roommates, four chill roommates. It's two boys and two girls. Everyone can invite one or two people and they will stay in the room and not in the kitchen. So, the possible infections remain very limited, so to speak*' (JT2005L2). Some agreements were made about physical contact and keeping distance to keep possible infections limited. This is common throughout the data, with students often explaining they made some sort of agreements on how to deal with the

COVID-19 restrictions. They also focused on other elements related to the COVID-19 regulations, such as the number of people.

In many student houses, COVID-19 restrictions were discussed, and intentions were expressed, but that did not necessarily result in very strict or clear agreements. A student who lives together with one other roommate answered the question of whether she had made agreements with her roommate, as follows:

*‘Hmm, not necessarily. We’ve always been pretty much the same: you don’t go to parties. I also know that there are people who have gone to house parties with really twenty people or whatever, or in the back of bars they had some kind of door to an illegal bar. And how nice that seemed to me. I just didn’t think that was possible and neither did he. So, we were just the same about that. (...) We were actually always the same about that. We haven’t made any verbal agreements or anything. Because we had the same idea’ (DR2406L2).*

Some agreements were made with family members, depending on how strictly parents or other family members followed the COVID-19 restrictions, or depending on their health status. Some students agreed with their parents that they would not visit if they were ill or had a cold. Later, when testing became available, they agreed to meet or to celebrate birthdays and holidays only after having tested negative. Students living with their parents made agreements concerning visiting friends and meeting them under certain conditions, such as only outside.

A recurring theme was the apparent interchangeability between adhering to and violating rules, with some students stating that if they adhered to a certain COVID-19 restriction, it would not be essential to follow others. The consensus appeared to be that it is unnecessary to follow all the rules and regulations, but rather to pick-and-choose from the presented options. As one of them explained:

*‘Yeah, officially you were only allowed to have two visitors at the time, and we were with four: me, my brother, and our partners. So, four people visiting, we violated the rules at the time. But we always kept enough distance, even when you needed to go the kitchen, those moments that you mindlessly ignore social distancing, that you don’t keep 1.5 meters distance. We also stood up one by one from the dining table’ (DR2406L2).*

### 5.3.3 Actual behavior

The data showed that intended behavior did not consistently align with actual behavior, with the most notable discrepancies in (not) adhering to COVID-19 restrictions observed in relation to 1) the number of people, 2) physical contact, and 3) keeping distance. Our investigation aimed to elucidate factors contributing to the discrepancy between intended and observed behavior. Whereas findings regarding intended behavior and considerations were mostly derived from the interviews, the results regarding actual behavior and context factors relied more heavily on both interviews and observations. The observations formed a useful tool for verifying the interview statements and triangulating some of the findings regarding actual behavior and the context in which it occurs.

Most behavioral deviation from intended behavior could be attributed to the contexts in which individuals found themselves, as will be elaborately discussed in the next section. For example, encountering other people presented a challenge: on the one hand individuals had an idea about what COVID-19 restrictions to adhere to, but on the other hand, reverting to previously ingrained “normal” behavior felt safe and familiar. Students noticed they needed to put effort into displaying their intended behavior when it deviated from their pre-COVID-19 habits.

Furthermore, a substantial portion of behavior was influenced by the tendency to mimic the actions of others. This phenomenon was evident in our observations, where students who initially adhered to distancing norms adjusted their behavior when new individuals joined their group. Case 6B2611BUJ3 provided a noteworthy illustration of this phenomenon as *‘a student walked with another student and initially maintained some distance while waiting. However, when a third person arrived, they exchanged greetings and made physical contact before continuing their walk together shoulder to shoulder’*.

Discrepancies between intended behavior and actual behavior also became clear in the interviews. Several students explained they did not keep distance or did meet with more people than was allowed due to their need for some intimacy with others. Students also explained that when they had been drinking it was more difficult to keep their distance: *‘(...) if I would be tipsy, then I would also definitely be that person having closer contact. So, suddenly, hugging people’* (DR2406L1). This is exemplary for how most students reasoned, having certain intentions that were discarded when the situation made it impossible or difficult to adhere to those intentions.

### 5.3.4 Context factors

The discrepancy between intended and actual behavior could be explained by the context of the situation. One major reason is the declining quality of the interaction with other people when the COVID-19 restrictions were adhered to. By keeping distance, avoiding physical contact, and interacting with fewer people than usual, the quality of social interactions decreased, either because being physically close expresses sympathy and intimacy or because keeping distance made it difficult to explain or demonstrate relevant aspects of the conversation. One participant illustrated this with an example of a moment where their large family was eating dinner at different tables: *‘(..) at one point my cousin joined us after the dinner, and we had such a nice conversation. If he would have stayed at his parents’ table, we would have never had this moment’* (JO0406L2).

One of the reasons presented earlier for not adhering to the COVID-19 restrictions was that students considered it harmful for their mental health and were in dire need of contact with others. At the same time, some students expressed that strict adherence to COVID-19 restrictions, such as maintaining a certain distance, was not desired by their friends and family either. As one student explained how her actual behavior around elderly people, over time, started to deviate from her intended behavior because of other people’s desires and behavior: *‘I often tried to keep my distance from older people and such from the start. And then you often get comments ‘that’s really not necessary, girl’. And then I also thought, wow, those people, (...) I think the older generation also thinks it’s really shitty. And yes, that was sad, so I didn’t do that [keeping distance] anymore’* (MB2705B1).

Furthermore, the physical environment often did not accommodate meeting according to COVID-19 restrictions. Though a location may meet the strict government-prescriptions, keeping distance in a somewhat crowded and noisy room makes it difficult to have a conversation, and sometimes the set-up of the room did not offer the needed space to move around. Such was the case for observation 3A2810PSJ103, where students sought shelter for the rain under the canopy of one of the food trucks, thereby failing to adhere to social distancing-regulations.

One of the students sketched the hectic and difficult physical situation when her student house was in quarantine:

*‘When we were in quarantine, it was complicated [to keep distance]. Not just because we had to do Tetris in the hallway, because someone had to go to the*

*toilet and the other had to go to the shower and yes, who goes first and where do you wait, but more mentally to be constantly alert that you don't come close to each other. Yes, that is very tiring (...)' (JT2005L3).*

The declining quality of the interaction, the physical environment, and the behavior of their contacts, served as reasons why students deviated from intended behavior. It shows how students often prioritized the contact with others over their initial ideas and their willingness to adjust their behavior away from their intentions.

It is noteworthy to mention that context factors also contributed to adhering to the COVID-19 regulations. Throughout the observations we noticed students cleaning their hands, following walking routes, sitting in assigned places, or looking at or discussing the interventions. Furthermore, students were actively approached by security guards and COVID-coordinators, people with the role to make sure everyone adheres to COVID-19 restrictions at the campus. Such as in the case 1J1210BI419, where a student walked shoulder to shoulder with another student, until the security guard made a remark about social distancing. They kept their distance afterwards and were laughing about the situation.

### 5.3.5 Construction of the empirical model

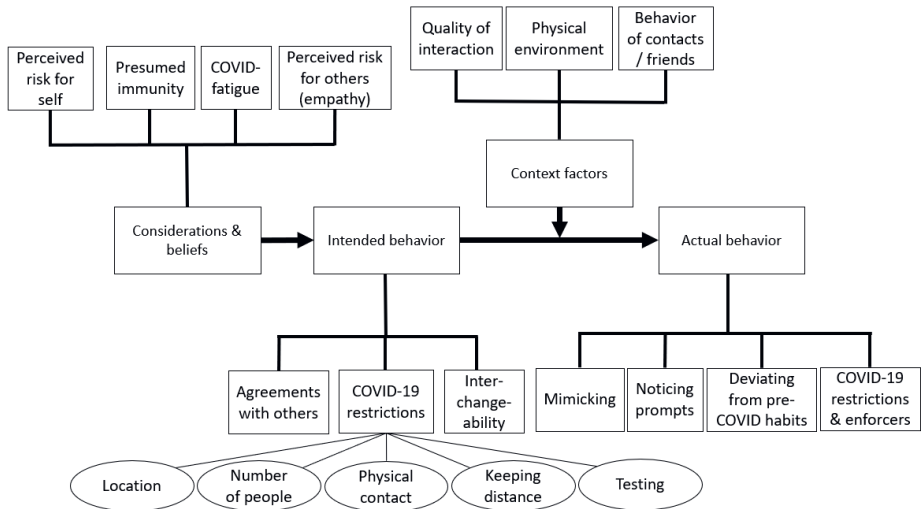
The results thus far, describing the process of going from considerations through intended behavior to actual behavior, including the moderation by the specific context of the social interaction, is visualized in Figure 5.5.

### 5.3.6 Comparison control versus experimental group

During the analysis, special attention was paid to empathy and the effects of the interventions. Although empathy played a role in the process from considerations, via intended to actual behavior, no clear evidence was found for the effectiveness of the prompts. Their main impact was attracting attention, serving as something to look at or talk about. However, behavior did not change in a meaningful way regarding compliance with COVID-19 restrictions. Minor differences were observed between control and experimental groups in their considerations, but these diminished when moving from considerations to actual behavior.

From the observation data we deduced that the direct effect of the prompts was that they drew attention. With moving visuals and sound, curiosity was expected. Yet, instead of changing their behavior, people usually looked at the

**Figure 5.5: The empirical model that visualizes how considerations lead to intentions and subsequently to behavior, moderated by contextual factors of an encounter.**



prompts and continued as before. The prompts occasionally served as a topic of conversation, visible through gestures and reactions, but rarely led to direct behavioral adjustments.

Regarding COVID-19 restrictions, two broad patterns were observed in both conditions. Some individuals tried to comply with rules such as distancing or following walking routes but could be persuaded to violate them when blocked by others or when security employees encouraged them to take shortcuts. Others, often in groups, ignored rules altogether, disregarding distancing or seating instructions. In these cases, prompts were ineffective, whereas interventions by security guards or COVID-coordinators did have immediate, if temporary, effects on compliance.

From the interviews some themes could be identified that were more present in either the control or experimental group. Students in the experimental group mentioned more frequently that: they considered the vulnerability of others (67% of cases versus 58% in control group;  $n = 4$  and  $n = 19$ ), they were less strict when being with people they are close with (50% versus 39%;  $n = 3$  and  $n = 13$ ), they considered low risk for themselves (50% versus 21%,  $n = 3$  and  $n = 7$ ), as well as COVID-fatigue (100% versus 81%;  $n = 6$  and  $n = 27$ ). Only the first suggests a possible positive influence of the prompts, as acknowledging

others' vulnerability reflects empathy. The control group, however, more often referred to adhering to restrictions deemed relevant by others (45% versus 0% in the experimental group;  $n = 15$  and  $n = 0$ ) or restrictions that are valued by vulnerable persons (36% versus 17%;  $n = 12$  and  $n = 1$ ).

When moving from considerations to intended and actual behavior, group differences diminished further. A difference in the actual behavior was that participants in the intervention group (17% of cases;  $n = 1$ ) more often described being inattentive to restrictions during interactions compared to the control group (6% of cases;  $n = 2$ ). Yet this small difference could just as easily be attributed to chance as to the prompts.

To further test the presence of an effect of the intervention, we tried to identify which group any random observation or interview was allocated to. Distinguishing whether interviews came from the control or intervention group proved to be difficult. Observation data showed clearer differences, largely because the prompts acted as distractions during data collection. However, regarding actual adherence to restrictions, knowing whether a person belonged to the control or intervention group did not improve the ability to predict behavior or intentions.

The overall outcome of this comparison revealed a lack of meaningful behavioral differences between control and experimental groups. Prompts successfully attracted attention and sometimes initiated discussion but did not substantially influence compliance with COVID-19 restrictions. Security staff interventions proved more effective in directly altering behavior, though typically only temporarily.

## 5.4 Discussion

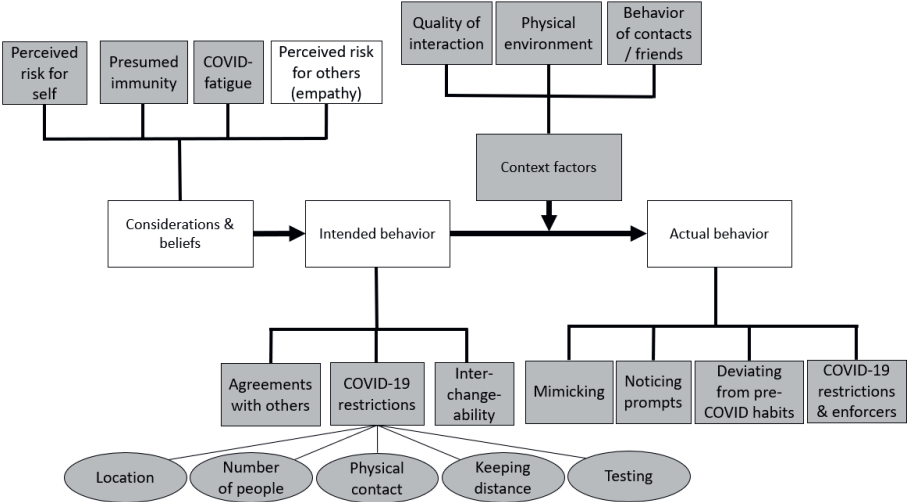
In this field experiment we researched how prompting affected behavioral outcomes in the context of COVID-19 restrictions. Because of the focus on both intended and actual behavior, the study relied on a combination of interview and observation data. These different data sources served as input for substantiating different aspects of the theoretical model.

From the interviews a model emerged explaining how behavior develops through considerations and beliefs, via intended behavior, to actual behavior within a certain context. This process appeared in all interviews, aligning with cognitive behavior theory (Beck, 1995). Prompts could have influenced behavior by shaping context or making considerations salient.

The interview data revealed some differences between the experimental and the control group in terms of considerations and intended behavior. However, no

differences between the groups in actual behavior were identified based on both interview and observation data. Moreover, the observation data made clear that the prompts drew people’s attention, but they did not act accordingly. Thus, the empirical model (Figure 5.5) deviates from the theoretical one (Figure 5.1). Figure 5.6 visualizes these differences: white boxes indicate overlap, grey boxes additions.

**Figure 5.6: Adjusted version of the empirical model to identify adjustments to the theoretical model.**



A few aspects of Figure 5.6 are noteworthy. First, the intervention is absent in this model, even though the original research question was about a comparison between the experimental conditions. The underlying reason is straightforward: no meaningful differences emerged between the groups, offering no clear evidence that the prompts had any effect. Second, only one aspect of empathy is present in this model, that is, the desire to uphold others’ well-being. The other aspects of empathy were absent in the data. However, new considerations and beliefs were identified that were not present in the theoretical model. Third, context factors such as environment, interaction quality, and others’ behavior had a larger impact than anticipated, creating contrasts between intended and actual behavior. De Ridder et al., 2022 similarly concluded that empathy prompts did not promote distancing. Our findings align with this and other studies (e.g. Favero & Pedersen, 1996).

From the literature there are some possible explanations why the empirical

data in our study deviated from the theoretical model. To begin with, we identified explanations known as the looking glass self-theory (Cooley, 2017) and the phenomenon of mirroring (Fogassi et al., 2005). First, it appeared that students expressed an extensive reasoning on how they would behave, and why this seemed reasonable to them. But when it came to their actual behavior, their own intentions were quickly abandoned to match the behavior of others around them. This is an example of mirroring. That does raise an interesting situation, since if virtually all people mirror, where does behavior ever originate from? Here the looking glass self-theory comes into play, which describes that people base their sense of self on how they perceive others to perceive them. In summary: people expect others to have expectations and judgments on them and base their behavior on these presumptions. In the context of our COVID-research: if someone assumes that others expect them to have physical contact, they will initiate the contact. They perceive their behavior as mirroring, when actually, they behave as they assume they are expected to behave. It is a self-enforcing circular reasoning.

In the context of COVID-19, earlier research on restrictions implemented during the pandemic, found particular reasons for young adults (among others) for not sticking to COVID-19 restrictions. Research on social distancing shows that several people, even when they intend to commit to social distancing, do not stay sufficiently far away from others because few are able to make an accurate judgment of distance (Yamamoto & Nightingale, 2023). These researchers also found that intended behavior deviated from actual behavior. Moreover, an exploratory study that investigated the effect of the first week of the COVID-19 pandemic lockdown on Dutch adolescents showed relatively lower levels of empathic concern (Van de Groep et al., 2020). The study indicated that adolescents became relatively more emotionally self-focused when experiencing lockdown and social distancing. Opportunities for prosocial actions declined, as adolescents had fewer opportunities to feel with and support others due to their limited opportunities to interact with people outside their household. The COVID-19 pandemic and subsequently implemented restrictions had major impact on youth well-being. Several studies indicate that COVID-19 restrictions led to higher levels of depression and anxiety among young people (e.g., Glowacz & Schmits, 2020). Findings from these studies also explain and underscore our findings that students who were ‘COVID-fatigued,’ were more reluctant to adhere to COVID-19 restrictions and longed for social contact with family, friends and peers.

A final explanation for the lack of effects from the prompts could be the

low number of interviewees in the experimental condition in Wave 2 (long interviews). As the data collection progressed, we did fear that it would prove impossible to draw meaningful conclusions from such a small sample. However, by looking at the data of the experimental and the control group together, we could identify the process that results in actual behavior. Furthermore, the data from the short interviews and the observations were rich enough to draw meaningful conclusions on the (lack of) difference in actual behavior between the conditions. As a result, an updated theoretical framework emerged from the empirical data.

In conclusion, in this study no relevant effect of the empathy-inducing prompts appeared in students' actual behavior. The results are based on a mixture of observations and interviews and throughout all data the lack of impact of the prompts was apparent, supporting the rigor of the results. We did recognize some behavioral processes that might explain why people behaved in a certain way and why the prompts were less influential. The concepts of mirroring and the looking glass self-theory might better explain the dominant strategies to form behavior than the prompts. Whether prompts are indeed ineffective in real life situations or have more impact in other aspects of reasoning is food for thought and serves as a starting point for future research.

## Bibliography

- Alkhalidi, G., Hamilton, F., Lau, R., Webster, R., Michie, S., & Murray, E. (2016). The effectiveness of prompts to promote engagement with digital interventions: A systematic review. *Journal of Medical Internet Research*, 18(5). <https://doi.org/10.2196/jmir.4790>
- Archibald, M. (2016). Investigator triangulation: A collaborative strategy with potential for mixed methods research. *Journal of Mixed Methods Research*, 10, 228–250. <https://doi.org/10.1177/1558689815570092>
- Beck, J. (1995). *Cognitive therapy: Basics and beyond*. New York: Guilford press.
- Brown, T., Ham, S., & Hughes, M. (2010). Picking up litter: An application of theory-based communication to influence tourist behavior in protected areas. *Journal of Sustainable Tourism*, 18(7), 879–900. <https://doi.org/10.1080/09669581003721281>
- Cohen, D., & Strayer, J. (1996). Empathy in conduct-disordered and comparison youth. *Developmental Psychology*, 32, 988–998. <https://doi.org/10.1037/00121649.32.6.988>
- Cooley, C. (2017). *Human nature and the social order*. Routledge. <https://doi.org/10.4324/9780203789513>
- Corbin, J., & Strauss, A. (1990). Grounded theory research: Procedures, canons and evaluative criteria. *Qualitative Sociology*, 13, 3–21. <https://doi.org/10.1007/BF00988593>
- Cuff, B., Brown, S., Taylor, L., & Howat, D. (2016). Empathy: A review of the concept. *Emotion Review*, 8(2), 144–153. <https://doi.org/10.1177/1754073914558466>
- De Leon, E., Fuentes, L., & Cohen, J. (2014). Characterizing periodic messaging interventions across health behaviors and media: Systematic review. *Journal of Medical Internet Research*, 16(3). <https://doi.org/10.2196/jmir.2837>
- De Ridder, D., Aarts, H., Benjamins, J., Glebbeek, M., Leplaa, H., Leseman, P., Potgieter, R., Tummers, L., & Zondervan-Zwijnenburg, M. (2022).

- “keep your distance for me”: A field experiment on empathy prompts to promote distancing during the covid-19 pandemic. *Journal of Community & Applied Social Psychology*, 32, 755–766. <https://doi.org/10.1002/casp.2593>
- Favero, N., & Pedersen, M. (1996). How to encourage “togetherness by keeping apart” amid covid-19? the ineffectiveness of prosocial and empathy appeals. *Journal of Behavioral Public Administration*, 3(1), 1–18. <https://doi.org/10.30636/jbpa.31.167>
- Fogassi, L., Ferrari, P., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science*, 308, 662–667. <https://doi.org/10.1126/science.1106138>
- Glaser, B. (1978). *Theoretical sensitivity*. Mill Valley, CA: The Sociology Press.
- Glowacz, F., & Schmits, E. (2020). Psychological distress during the covid-19 lockdown: The young adults most at risk. *Psychiatry Research*, 291. <https://doi.org/10.1016/j.psychres.2020.113486>
- Grant, A., & Hofmann, D. (2011). It’s not all about me: Motivating hand hygiene among health care professionals by focusing on patients. *Psychological Science*, 22, 1494–1499. <https://doi.org/10.1177/0956797611419172>
- Gutiérrez-Cobo, M., Cabello, R., Megías-Robles, A., Gómez-Leal, S., R. and Baron-Cohen, & Fernández-Berrocal, P. (2023). Does our cognitive empathy diminish with age? the moderator role of educational level. *International Psychogeriatrics*, 35(4), 207–214. <https://doi.org/10.1017/S1041610221000624>
- Lehman, P., & Geller, E. (2004). Behavior analysis and environmental protection: Accomplishments and potential for more. *Behavior and Social Issues*, 13, 13–33. <https://doi.org/10.5210/bsi.v13i1.33>
- Leoniak, K., & Cwalina, W. (2019). The role of normative prompts and norm support cues in promoting light-switching behavior: A field study. *Journal of Environmental Psychology*, 64, 1–11. <https://doi.org/10.1016/j.jenvp.2019.04.014>
- Lunn, P., Timmons, S., Belton, C., Barjaková, M., Julianne, H., & Lavin, C. (2019). Motivating social distancing during the covid-19 pandemic: An online experiment. *Social Science Medicine*, 265, 1113478. <https://doi.org/10.1016/j.socscimed.2020.113478>
- Malin, H., Reilly, T., Quinn, B., & Moran, S. (2014). Adolescent purpose development: Exploring empathy, discovering roles, shifting priorities, and creating pathways. *Journal of Research on Adolescence*, 24(1), 186–199. <https://doi.org/10.1111/jora.12051>

- Maltinsky, W., & Swanson, V. (2020). Behavior change in diabetes practitioners: An intervention using motivation, action planning and prompts. *Patient Education and Counseling*, 103(11), 2312–2319. <https://doi.org/10.1016/j.pec.2020.04.013>
- Marci, C., Ham, J., Moran, E., & Orr, S. (2007). Physiologic correlates of perceived therapist empathy and social-emotional process during psychotherapy. *The Journal of Nervous and Mental Disease*, 195(2), 103–111. <https://doi.org/10.1097/01.nmd.0000253731.71025.fc>
- Martinez, A., & Knee, R. (2024). Using self-determination theory to link empathy and voluntary health behaviors. *Journal of Applied Social Psychology*, 1–12. <https://doi.org/10.1111/jasp.13064>
- McMullin, C. (2023). Transcription and qualitative methods: Implications for third sector research. *Voluntas*, 34, 140–153. <https://doi.org/10.1007/s11266-021-00400-3>
- Oliveira-Silva, P., & Gonçalves, O. (2011). Responding empathically: A question of heart, not a question of skin. *Applied psychophysiology and biofeedback*, 36, 201–207. <https://doi.org/10.1007/s10484-011-9161-2>
- Patton, M. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research*, 34, 1189–1208. <https://doi.org/PMC1089059>
- Pfattheicher, S., Nockur, L., Böhm, R., Sassenrath, C., & Petersen, M. B. (2020). The emotional path to action: Empathy promotes physical distancing and wearing of face masks during the covid-19 pandemic. *Psychological Science*, 31(11), 1363–1373. <https://doi.org/10.1177/0956797620964422>
- Prinsen, S., de Ridder, D., & de Vet, E. (2013). Eating by example. effects of environmental cues on dietary decisions. *Appetite*, 70, 1–5. <https://doi.org/10.1016/j.appet.2013.05.023>
- QSR International Pty Ltd. (2020). Nvivo [Available from <https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home>].
- Schieman, S., & Van Gundy, K. (2000). The personal and social links between age and self-reported empathy. *Social Psychology Quarterly*, 63(2), 152–174. <https://doi.org/10.2307/2695889>
- Van de Groep, S., Zanolie, K., Green, K., Sweijen, S., & Crone, E. (2020). A daily diary study on adolescents' mood, empathy, and prosocial behavior during the covid-19 pandemic. *PloS One*, 15(10), 1–20. <https://doi.org/10.1371/journal.pone.0240349>

- Whitehead, M., Jones, R., Howell, R., Lilley, R., & Pykett, J. (2014). Nudging all over the world: Assessing the global impact of the behavioural sciences on public policy. <https://api.semanticscholar.org/CorpusID:197770852>
- Wichmann, B., Luckert, M., Bissonnette, K., Cumberland, A., Doll, C., Gupta, T., & Shi, Y. (2010). Using pro-environmental information to modify conservation behavior: Paper recycling and reuse. *Recycling*, 2(1), 5. <https://doi.org/10.3390/recycling2010005>
- Yamamoto, N., & Nightingale, M. (2023). How well do we do social distancing? *Quarterly Journal of Experimental Psychology*, 77(5), 1106–1112. <https://doi.org/10.1177/17470218231195247>

## Appendix 5.A: Observation list control week

**Appendix 5.A: Observation list control week.** A Dutch version was used during data collection, with a table per half hour. We offer the translated example for readability.

We filled in the observation list to show an example of what kind of data was observed. In this example, the researcher observed from 10 to 10.30h, and saw a total of 11-20 people in that timeframe. They conducted four observations, e.g. Person 1 who actively avoided some people, but did not continuously keep 1.5m distance, or Person 3 wears a face mask and does not encounter others (therefor neither does nor doesn't keep distance). Field notes, rightmost column, give more details on the observed people.

Table 5.2: *Example of filled in observation list, not actual data:*

|                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                 |                 |                     |                                                          |                                               |                     |                        |                                   |
|---------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|-----------------|---------------------|----------------------------------------------------------|-----------------------------------------------|---------------------|------------------------|-----------------------------------|
| Name observant                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                 |                 |                     |                                                          |                                               |                     |                        |                                   |
| Location observation:                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                 |                 |                     |                                                          |                                               |                     |                        |                                   |
| <input type="checkbox"/> square outside a college hall        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                 |                 |                     | <input type="checkbox"/> main entrance of a lecture hall |                                               |                     |                        |                                   |
| <input checked="" type="checkbox"/> entrance to lecture rooms |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                 |                 |                     |                                                          |                                               |                     |                        |                                   |
| Time period                                                   | Number of people at the location                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Person observed | Wears face mask | Keeps 1.5m distance | Addresses others to keep their distance                  | Avoids people actively to keep their distance | Uses hand sanitizer | Does not keep distance | Does not keep distance to friends |
| 10:00-10:30                                                   | <input type="checkbox"/> 0-10                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 1               | 1               | 1                   |                                                          | 1                                             |                     |                        |                                   |
|                                                               | <input checked="" type="checkbox"/> 11-20                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 2               | 2               |                     |                                                          |                                               | 2                   |                        | 2                                 |
|                                                               | <input type="checkbox"/> 21-30                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | 3               | 3               |                     |                                                          |                                               |                     |                        |                                   |
|                                                               | <input type="checkbox"/> 31-40                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | 4               |                 |                     |                                                          |                                               |                     | 4                      | 4                                 |
|                                                               | <input type="checkbox"/> 41-50                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |                 |                 |                     |                                                          |                                               |                     |                        |                                   |
|                                                               | <input type="checkbox"/> 50+                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |                 |                 |                     |                                                          |                                               |                     |                        |                                   |
| Field notes                                                   | <div>1. F, walks in and runs into two people who are standing around talking. Waits for that before walking on.</div> <div>2. F, walks in with girlfriend. Have both put on mouthpiece before entering. Uses alcohol dispenser and keeps distance when girlfriend uses it. Then walks in walking direction upstairs. Otherwise few people in the room.</div> <div>3. M, is alone. Wears mouth mask and uses alcohol dispenser. Walks through to the cafeteria.</div> <div>4. M, is with a colleague. Does not wear mouth mask over nose. Colleague does wear mouth mask properly. Does not keep distance from each other while walking. Moves a little further apart when they see the security guard.</div> |                 |                 |                     |                                                          |                                               |                     |                        |                                   |

## Appendix 5.B: Observation list intervention week

**Appendix 5.B: Observation list intervention week.** A Dutch version was used during data collection, with a table per half hour. We offer the translated example for readability.

We filled in the observation list to show an example of what kind of data was observed. In this example, the researcher observed from 8.30 to 9h, and saw a total of 21-30 people in that timeframe. They conducted three observations, e.g. Person 1 who actively avoided some people, but did not continuously keep 1.5m distance, or Person 2 who did not keep their distance at all. Field notes, rightmost column, give more details on the observed people.

Table 5.3: *Example of filled in observation list, not actual data:*

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                  |                                           |                 |                     |                                         |                                               |                     |                        |                                   |                                     |                       |                                       |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------|-------------------------------------------|-----------------|---------------------|-----------------------------------------|-----------------------------------------------|---------------------|------------------------|-----------------------------------|-------------------------------------|-----------------------|---------------------------------------|
| Name observant                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |                                  |                                           |                 |                     |                                         |                                               |                     |                        |                                   |                                     |                       |                                       |
| Location observation:                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |                                  |                                           |                 |                     |                                         |                                               |                     |                        |                                   |                                     |                       |                                       |
| <input checked="" type="checkbox"/> square outside a college hall <input type="checkbox"/> main entrance of a lecture hall <input type="checkbox"/> entrance to lecture rooms                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                  |                                           |                 |                     |                                         |                                               |                     |                        |                                   |                                     |                       |                                       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Number of people at the location | Person observed                           | Wears face mask | Keeps 1.5m distance | Addresses others to keep their distance | Avoids people actively to keep their distance | Uses hand sanitizer | Does not keep distance | Does not keep distance to friends | Does not keep distance to strangers | Notices interventions | Stands still in front of intervention |
| Time period                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                  |                                           |                 |                     |                                         |                                               |                     |                        |                                   |                                     | Looks at intervention | Discusses intervention with others    |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 10:00-10:30                      | <input type="checkbox"/> 0-10             | 1               |                     |                                         | 1                                             | 1                   |                        |                                   | 1                                   |                       |                                       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                  | <input type="checkbox"/> 11-20            | 2               |                     |                                         |                                               | 2                   | 2                      |                                   | 2                                   |                       |                                       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                  | <input checked="" type="checkbox"/> 21-30 | 3               |                     |                                         | 3                                             | 3                   |                        |                                   | 3                                   |                       |                                       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                  | <input type="checkbox"/> 31-40            |                 |                     |                                         |                                               |                     |                        |                                   |                                     |                       |                                       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                  | <input type="checkbox"/> 41-50            |                 |                     |                                         |                                               |                     |                        |                                   |                                     |                       |                                       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                  | <input type="checkbox"/> 50+              |                 |                     |                                         |                                               |                     |                        |                                   |                                     |                       |                                       |
| Field notes                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                  |                                           |                 |                     |                                         |                                               |                     |                        |                                   |                                     |                       |                                       |
| 1. F, was alone, crosses the square. there are 2 long rows at the food stalls, at one she walks around it with a short turn. The second row extends to the library so she cannot go around it and has to go through it. She quickens her stride as she crosses through the row and bows her head down, the back man in the row takes a small step back to give her space but still they do not keep 1.5m distance.                                                                                                                                                                                                                                                                     |                                  |                                           |                 |                     |                                         |                                               |                     |                        |                                   |                                     |                       |                                       |
| 2. F, arrives with friend, they cross the square, go through 1 of the rows in front of the food stalls to get to the second one and join the back. She walks and stands close to her friend while they are busy talking. She does not keep 1.5m distance from the person in front of her in the row. When people cross the square she makes no effort to give them space to pass through. Another 2 friends join her in line, again she does not keep a distance from them.                                                                                                                                                                                                            |                                  |                                           |                 |                     |                                         |                                               |                     |                        |                                   |                                     |                       |                                       |
| 3. M was alone, approaches from the library and gets in line at the first food stand. He keeps about a meter distance from the person in front of him in line. When a group of people arrive and want to pass behind him he takes a small step forward, he keeps stepping back and forth, doesn't seem to know quite where to go as he is now closer to the person in front of him. Takes another step backwards when they have passed. When someone gets in line behind him he seems to notice this at first and feel uncomfortable about it but after a while he seems to have forgotten this and stays still and looks ahead. When he walks away he does not encounter anyone else. |                                  |                                           |                 |                     |                                         |                                               |                     |                        |                                   |                                     |                       |                                       |

## Appendix 5.C: Topic list short interviews

**Appendix 5.C: Topic list short interviews.** A Dutch version of the topic list was used during short interviews, but we offer a translated version for readability.

Introduction: Hi! Can I ask you two question for a research on the COVID-19 restrictions? It will take 5 to 10 minutes, and I can walk along with you during the interview. We do audio-record the interview, for analysis at a later time.

Thank you for agreeing to participate. Please answer the questions from the perspective of your normal day-to-day life, the questions are not focused on the situation at Utrecht Science Park.

I start the audio-recording now.

First some background questions: what study program are you a part of? In which year of the program?

- Question 1. What do you think of the COVID-19 restrictions?
- Question 2. Do you intend to adhere to the COVID-19 restrictions?

In the next few weeks some more extended, longer interviews, will be conducted in a project related to this one. Would you like to participate? This interview can be conducted at a time that suits you, face-to-face or via Teams.

I have one final question for you: How would you feel if you were being observed in this time of crisis in the context of a COVID-19 related research? Possible explanation: for instance a study that would document behavior in public spaces, in order to determine how well the COVID-19 restrictions are adhered to, and to investigate how people would react to aspects related to the restrictions.

Potential follow-up questions to ask during the interview:

- Question for clarification, for example:
  - Can you explain that?

- What do you mean with ...?
  - How should I interpret what you just said?
  - In what way do you mean this?
- Text for elaboration:
  - With the COVID-19 restrictions we refer to the restrictions that the government takes to reduce spreading the COVID-virus. This includes the call to keep 1.5m distance, to wear face-masks in public spaces, and to limit social contact.
- Follow-up questions:
  - And how about other situations? [if the participant has a big focus on situations outside the university]

## Appendix 5.D: Topic list long interviews

**Appendix 5.D: Topic list long interviews.** We used a Dutch version during the long interviews, but offer a translated version for readability.

| Topic subtopics   | with | Check | Questions & Follow-up questions                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|-------------------|------|-------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Start             |      |       | Give explanation over informed consent and let them sign the form (if they haven't already).                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Informed consent  |      |       | <p>Acquire relevant demographic information through informal conversation (gender, age, study program, year of study)</p> <p>Give explanation regarding study: During this interview we will talk about life in times of the COVID-19 crisis. To structure the conversation we have selected some subtopics based on which we can discuss this topic. We are very much interested in the way you, as a student, experience this situation. You are encouraged to talk from your own perspective, and provide information on the context of the situation if that might be relevant.</p> |
| Studying          |      |       | How is your study currently organized?                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| Ability to adjust |      |       | <p>Follow up on:</p> <ul style="list-style-type: none"><li>• Physical education,</li><li>• amount of contact with teachers,</li><li>• amount of contact with fellow students.</li></ul> <p>What do you think of studying in this form?</p> <ul style="list-style-type: none"><li>• Follow up on:</li><li>• Is the study going okay?</li><li>• Did it require adjustment time?</li></ul>                                                                                                                                                                                                 |

|                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|-----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Attitude                          | <p>What are your ideas regarding the governments COVID-19 restrictions?</p> <ul style="list-style-type: none"><li>• How do you feel about the amount of curbing of the restrictions?</li><li>• Are the restrictions in proportion to the situation?</li><li>• If new restrictions are announced, how you know?</li><li>• Are you actively trying to stay up-to-date of all restrictions and developments?</li><li>• Why do feel this is (not) important to do?</li></ul> |
| Daily life                        | <p>Can you explain how your week looks like?</p> <ul style="list-style-type: none"><li>• What do you do at home?</li><li>• When and why do you leave home?</li><li>• How do you make decisions about this?</li></ul>                                                                                                                                                                                                                                                     |
| Social life                       | <p>Can you describe how your social life looks before the COVID-19 restrictions?</p>                                                                                                                                                                                                                                                                                                                                                                                     |
| Study-related contact             | <p>What does having contact with other students or teachers look like?</p> <ul style="list-style-type: none"><li>• How did you arrive at this form?</li><li>• How do you think others perceive this?</li><li>• How do other students have contact?</li><li>• How do you feel about this?</li></ul>                                                                                                                                                                       |
| Attitude regarding social contact | <p>Do you experience your social life differently now than before?</p> <ul style="list-style-type: none"><li>• In what way?</li><li>• How do you feel about this?</li><li>• If not, do you feel the COVID-19 restrictions affect your contact moments in a different way?</li></ul>                                                                                                                                                                                      |
| Friends                           | <p>What does having contact with your friend look like currently?</p> <ul style="list-style-type: none"><li>• How do you feel about this?</li><li>• How do you make decisions about this?</li></ul>                                                                                                                                                                                                                                                                      |

|                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Home situation                           | <p>Can you describe your living situation?</p> <ul style="list-style-type: none"> <li>• With whom are you living there?</li> <li>• How do you experience the current situation?</li> <li>• Follow up on:             <ul style="list-style-type: none"> <li>– How do you deal with the COVID-19 restrictions?</li> <li>– Is your living situation changed?</li> <li>– Have you made certain agreements?</li> </ul> </li> </ul> <p>If living in a student house, follow up on:</p> <ul style="list-style-type: none"> <li>• Do you visit your parent's home (i.e. "homehome")</li> <li>• Do you deal with this differently than before?</li> <li>• Did you make any agreements with them?</li> </ul> |
| Family                                   | <p>If living at their parent's home, follow up on:</p> <ul style="list-style-type: none"> <li>• When, and what for, do you leave home?</li> <li>• How do other family members do this?</li> <li>• Did you make any agreements with them?</li> <li>• Why (not)?</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                           |
| Study/student/<br>sport-<br>associations | <p>Are you currently active in a study-, student, or sport association?</p> <ul style="list-style-type: none"> <li>• How is that currently organized?</li> <li>• How do you feel about this?</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Work                                     | <p>Do you have a job next to your study at the moment?</p> <p>If yes,</p> <ul style="list-style-type: none"> <li>• Do you have contact with other people in your job?</li> <li>• If ja, how do you feel about this?</li> <li>• If no, was this different before the COVID-19 restrictions?</li> <li>• How do you feel about this?</li> </ul> <p>If no,</p> <ul style="list-style-type: none"> <li>• Did you have a job next to your study before COVID-19?</li> </ul>                                                                                                                                                                                                                               |

|                         |  |                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|-------------------------|--|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Mood / State<br>of mind |  | <p>Do you perceive the current “new normal” to have an effect on your state of mind?</p> <ul style="list-style-type: none"><li>• If yes, how has it changed?</li><li>• If no, do you think other students around your perceive an effect?</li></ul> <p>Do you feel that you now get happiness from other things than before the measures were announced?</p> <ul style="list-style-type: none"><li>• If yes, in what way?</li></ul> |
| Future                  |  | <p>What have you learned from this period?</p> <ul style="list-style-type: none"><li>• Are there certain things you want to continue doing when the COVID-19 pandemic is over?</li><li>• What are moments that give you a lot of happiness?</li></ul> <p>What do you look most forward to once the COVID-19 pandemic is over?</p> <ul style="list-style-type: none"><li>• What would be the first thing you would do?</li></ul>     |
| Wrapping up             |  | <p>Are there things left unsaid during this interview about this topic that should be discussed?</p>                                                                                                                                                                                                                                                                                                                                |
| Final<br>question       |  | <p>How would you feel if you were being observed in public spaces such as a train station, in a supermarket, or in other public buildings, in this time of crisis in the context of a COVID-19 related research?</p> <p>Follow up on:</p> <ul style="list-style-type: none"><li>• Why would you feel this way?</li><li>• What do you think other people feel about being observed?</li></ul>                                        |

## CHAPTER 6

### In conclusion

This PhD-trajectory started with a plan of a line of research focused on replication by documenting the perspectives of researchers in psychology at the moment and by offering guidelines for a novel way of analyzing replication studies. In 2020, the onset of the COVID pandemic introduced disruptions in all aspects of life. At the same time, it created a valuable opportunity: I could join a field experiment related to COVID-related restrictions to explore the benefits of qualitative methods in experimental psychological research. Qualitative methods are a personal interest of mine, and the potential added value they have for psychological experimental research is something that I considered worth exploring. As a result, the pandemic heavily shaped this dissertation, leading to a structure divided into two parts: two chapters on replication studies conducted prior to COVID, and two chapters focusing on the use of qualitative methods in experimental research conducted during the pandemic. This division also reflects the passage of time between the two parts, with the first half of the dissertation representing work from about 6 years ago. Therefore, this concluding chapter discusses some recent developments in replication research.

First of all, there has been a slight increase in the presence of replication studies as well as an increase of research on replication in the literature. Torka et al. (2023) counted publications mentioning 'replicat' in the title, abstract, keywords, or author-defined keywords. This inquiry showed a growth of academic discourse around replication when comparing 2016-2022 with the period before 2016 (Torka et al., 2023). Clarke et al. (2024) studied the presence of direct replications in psychology since 2010. They identified studies as direct replication based on the screening of the abstract and, if required, a further examination of the contents of the study. Of the journals analyzed, very few published ten or more direct replications and more than half did not publish

any. The majority of direct replications they did find were published in 2019 or later. Their conclusion was that despite a slight increase in recent years, the number of direct replications is still small. Similar conclusions have been drawn in economics (Ankel-Peters et al., 2023; Sukhtankar, 2017) and education (Cook et al., 2025).

Also, there are indications that policies have shifted to push for more replication research. For example, the Dutch Research Council (NWO) established a dedicated funding scheme for replication studies between 2016 and 2022 (NWO, 2016). More recently, Open Science NL introduced a similar initiative in 2025 (OpenScienceNL, 2025). However, these efforts have been temporary in nature and there is not a sustainable institutional commitment to promoting replication research.

In 2025, a new study into the views on replication studies was published. The Replication in Action team investigated the perspective of researchers throughout the Netherlands (Derksen et al., 2025), as we did in 2019 from the perspective of researchers in psychology. In this paper the experiences of the 25 authors of the paper were combined with data from 68 interviews. The most common perspective closely aligned with the views of the *collectivist* in Chapter 2. This perspective states a preference for conceptual replications, stressed the need for transparency, and saw many benefits of replication studies. The Replication in Action team does not claim to form a representation of the general population or the most prevalent opinions within science. Rather, they acknowledge that their sample consists of researchers who have conducted a replication study, e.g., through funds from the NWO. It could be informative to explore the views of the broader population of researchers in the social sciences.

After studying the views of psychologists, we explored the role of Bayesian statistics in the context of replication studies. It was a response to the state of replication research in the social sciences at the time, as well as to the ways in which researchers commonly approach and interpret replication results. The method developed and presented in Chapter 3 offers researchers conducting replication studies alternative approaches to assessing whether the findings are (in)consistent with those of the original study.

The academic community has continued to explore the role Bayesian statistics can have in replication research. For instance, multiple replication studies have been published that apply Bayesian statistical methods to determine corroboration (e.g., Cutter et al., 2022; Gurunandan et al., 2023; Mera et al., 2025). Furthermore, some methodological studies have investigated the use of Bayesian statistics for sample size determination in the context of replication (e.g., Pawel

et al., 2023; Anderson & Kelley, 2024).

This small selection of recently published papers on replication shows the ongoing interest in the topic. My personal hope is that the presented work on qualitatively analyzed experiments sparks a similar interest. With this dissertation, I aim to add a small contribution to the continuous development of research methods and data analysis approaches in the social sciences. Who knows, maybe in some years the dreams of a certain high school student about a world with perfect science, scientists, and scientific research do come true?

## Bibliography

- Anderson, S. F., & Kelley, K. (2024). Sample size planning for replication studies: The devil is in the design. *Psychological methods*, 29(5), 844–867. <https://doi.org/10.1037/met0000520>
- Ankel-Peters, J., Fiala, N., & Neubauer, F. (2023). Do economists replicate? *Journal of Economic Behavior & Organization*, 212, 219–232. <https://doi.org/https://doi.org/10.1016/j.jebo.2023.05.009>
- Clarke, B., Lee, P. Y. K., Schiavone, S. R., Rhemtulla, M., & Vazire, S. (2024). The prevalence of direct replication articles in top-ranking psychology journals. *American Psychologist*. <https://doi.org/10.1037/amp0001385>
- Cook, B. G., Therrien, W. J., Waterfield, D. A., McClain, S., Fleming, J. I., Robinson, H., Watson, L., & Boyle, J. (2025). Same as it ever was: An updated review of replication studies in special education journals. *Remedial and Special Education*, 46(3), 236–248. <https://doi.org/10.1177/07419325241248766>
- Cutter, M. G., Filik, R., & Paterson, K. B. (2022). Do readers maintain word-level uncertainty during reading? a pre-registered replication study. *Journal of Memory and Language*, 125, 104336. <https://doi.org/10.1016/j.jml.2022.104336>
- Derksen, M., Meirmans, S., Brenninkmeijer, J., Pols, J., de Boer, A., van Eyghen, H., Gayet, S., Groenwold, R., Hernaus, D., Huijnen, P., Jonker, N., de Kleijn, R., Kroll, C. F., Kryptos, A.-M., van der Laan, N., Luijken, K., Meijer, E., Pear, R. S. A., Peels, R., ... de Winter, J. (2025). Replication studies in the netherlands: Lessons learned and recommendations for funders, publishers and editors, and universities. *Accountability in Research*, 32(7), 1285–1303. <https://doi.org/10.1080/08989621.2024.2383349>
- Gurunandan, K., Cooper, E., Tibon, R., Henson, R. N., & Greve, A. (2023). No evidence of fast mapping in healthy adults using an implicit memory measure: Failures to replicate the lexical competition results of

- coutanche and thompson-schill (2014). *Memory*, 31(10), 1320–1339. <https://doi.org/10.1080/09658211.2023.2262188>
- Mera, Y., Modirrousta-Galian, A., Thomas, G., Higham, P. A., & Seabrooke, T. (2025). Erring on the side of caution: Two failures to replicate the derring effect. *Journal of Experimental Psychology: General*, 3(154), 658–671. <https://doi.org/10.1037/xge0001707>
- NWO. (2016). Replicatiestudies pilot by NWO [Accessed: 2025-07-10 at <https://www.nwo.nl/onderzoeksprogrammas/replicatiestudies>].
- OpenScienceNL. (2025). Replication studies by open science nl [Accessed: 2025-07-10 at <https://www.openscience.nl/en/news/new-call-replication-research-becomes-more-visible-and-normalised>].
- Pawel, S., Consonni, G., & Held, L. (2023). Bayesian approaches to designing replication studies. *Psychological Methods*, 31(1), 22–39. <https://doi.org/10.1037/met0000604>
- Sukhtankar, S. (2017). Replications in development economics. *American Economic Review*, 107(5), 32–36. <https://doi.org/10.1257/aer.p20171120>
- Torka, A.-K., Mazei, J., & Hüffmeier, J. (2023). Are replications mainstream now? a comparison of support for replications expressed in the policies of social psychology journals in 2015 and 2022. *Credibility in Social and Personality Psychology*, 18, 1–22. <https://doi.org/10.32872/spb.9695>



## Nederlandse samenvatting

Aan de start van dit proefschrift was een onderzoekslijn gepland over replicatieonderzoek. COVID ontregelde, naast de rest van de wereld, ook dit plan. Door COVID kwam de kans aan te sluiten bij een veldexperiment naar gedragsbeïnvloeding. Het leidt tot een proefschrift met twee delen, maar wel binnen een overkoepelend doel. Het proefschrift is erop gericht een bijdrage te leveren aan de kwaliteit van de sociale wetenschappen door gebruikelijke methoden toe te passen in minder gebruikelijke contexten. Dat gebeurt eerst door in te gaan op stand van zaken met betrekking tot replicatieonderzoek, vervolgens door de toegevoegde waarde van kwalitatieve methoden in experimentele contexten te onderzoeken.

Hoofdstuk 2 gaat in op de replicatiecrisis in psychologie. Publicaties van het Open Science Collaboration toonden dat slechts 36 tot 47% van de studies in psychologie bij herhaling dezelfde conclusies zouden opleveren. In de jaren na deze publicaties is er een aanzienlijke hoeveelheid literatuur gepubliceerd waarin wordt ingegaan op replicatiestudies, hoofdzakelijk geschreven door methodologen en statistici. Er is echter weinig literatuur te vinden waarin de perspectieven van de onderzoekers binnen psychologie zelf ten opzichte van replicatie centraal worden gesteld. Met deze studie worden de meningen en ideeën van de onderzoekers binnen psychologie vastgelegd, op basis van 11 interviews, in 2019 uitgevoerd, met onderzoekers aan de afdeling Psychologie van een Nederlandse universiteit. Deze onderzoekers werden bevraagd over hun ideeën over waarom, hoe, en wanneer replicatiestudies zouden moeten worden uitgevoerd.

De data hebben geleid tot de ontwikkeling van een typologie waarin vier verschillende perspectieven ten aanzien van replicatiestudies beschreven worden. Deze typen worden beschreven op basis van tien karakteristieken, zoals ideeën over het doel van replicaties, voorkeuren voor methodologische benaderingen, de volgens hen gepaste wijze van analyse, evenals bredere concepten zoals hun ideeën over wetenschap en de uitdagingen die men daarin tegenkomt. De uiteindelijke typologie geeft een aantal hoofdthema's weer die ook in de eerdere literatuur al naar voren kwamen, maar geeft daarnaast ook nieuwe inzichten over hoe onderzoekers binnen de psychologie replicatie zien en ervaren.

Hoofdstuk 3 gaat door op het onderwerp replicatieonderzoek. Hierin wordt een methode voorgesteld om vast te stellen of de resultaten van een oorspronkelijke studie opnieuw worden gevonden in de replicatiestudie. Het paper wordt geïllustreerd met twee replicatiestudies en hun bijbehorende oorspronkelijke

---

studies, beide afkomstig uit het *reproducibility project psychology* van het Open Science Collaboration. In de voorgestelde methode is het opstellen van de te toetsen hypothesen flexibel, juist omdat het vertalen van 'het resultaat van een studie' naar hypothesen cruciaal zijn voor replicatie. De ene manier van het vaststellen van de hypothesen is door deze te formuleren op basis van de introductiesectie van de originele studie, wat leidt tot zogenaamde informatieve hypothesen. Alternatief kan er gekozen worden om de statistieken uit de resultatensectie van de originele studie om te zetten naar interval-hypothesen. Middels de Bayes factor kan dan worden vastgesteld of de hypothesen gebaseerd op de originele studie als meest waarschijnlijk naar voren komen op basis van de data in de replicatie. Met de voorgestelde methode om replicatiesucces te bepalen kunnen replicerende onderzoekers meer accuraat toetaen in hoeverre de studies vergelijkbare resultaten vinden.

Hoofdstuk 4 gaat in op het bepalen van effecten van interventies in een experimenteel design met behulp van kwalitatieve data en analysemethoden. Er worden verschillende stappen en afwegingen gepresenteerd die gemaakt zouden moeten worden bij de kwalitatieve evaluatie van resultaten van experimentele designs. Voor elk van de stappen wordt methodologische instructie geboden, van het formuleren van de onderzoeksdoelen, via data-verzameling en analyse, tot het interpreteren van het potentiële effect van de interventie. Daarbij wordt specifiek ingegaan op het versterken van de kwaliteit van de studie. De instructies zijn ontwikkeld aan de hand van een empirisch voorbeeld, dat in het paper als illustratie wordt gebruikt. In het empirische voorbeeld werd de constructivistische grounded theory benadering toegepast om te evalueren of er een effect is van prompts gericht op empathie, op gedrag en intenties met betrekking tot het opvolgen van COVID-gerelateerde richtlijnen. Middels de gepresenteerde stappen is het mogelijk voor andere onderzoekers om hun eigen kwalitatieve analyse van data afkomstig uit een experimenteel design uit te voeren.

Hoofdstuk 5 beschrijft de empirische studie die ten grondslag lag aan het vorige hoofdstuk. Het beschrijft het kwalitatieve experiment dat tijdens de COVID-pandemie werd uitgevoerd op de campus van een Nederlandse universiteit en hogeschool om te onderzoeken hoe empathie-oproepende prompts gedrag beïnvloeden. Middels een A-B design, wisselend tussen controle (A) en experimentele (B) weken, werd het effect van prompts op gedrag onderzocht. De prompts werden aangeboden in de vorm van video's, posters, en een robot. Deze prompts riepen op de COVID-richtlijnen op te volgen in het belang van anderen: 'Houd afstand voor mij'. De bevindingen tonen de complexiteit van het gedrag van de studenten, waar attitudes met betrekking tot COVID-gerelateerde richtlijnen invloed lijken te hebben op voorgenomen gedrag. Deze intenties leiden echter niet per se tot zichtbare veranderingen in het gedrag. De contextuele factoren waarin het gedrag tot stand kwam bleek de relatie daartussen te beïnvloeden. Daarnaast werden er slechts minimale verschillen vastgesteld tussen de controlegroep en de experimentele groep, wat suggereert dat de invloed van de prompts op gedrag zeer beperkt was.

Deze dissertatie is gestart voordat COVID de kop opstak. In de eerste jaren, voor COVID, zijn de studies uitgevoerd gericht op replicatie, tijdens COVID is

---

de focus verlegd naar kwalitatieve methoden. In Hoofdstuk 6 wordt een blik geworpen op de recente ontwikkelingen op het gebied van replicatie. In de jaren sinds Hoofdstuk 2 en 3 is het veld van replicaties in beweging gebleven. Ten eerste zijn er meerdere methodologische papers geschreven over de wijze waarop Bayesiaanse statistiek toegepast kan worden in de context van replicaties (e.g. om te bepalen of een studie succesvol gerepliceerd is of om de steekproefgrootte te bepalen). Ten tweede zijn er meer (directe) replicaties gepubliceerd, hoewel de toename gering genoemd kan worden. Het zijn indicaties van een wetenschappelijke wereld die in beweging is, en waarschijnlijk zal blijven. De hoop is dat deze dissertatie een bijdrage kan leveren aan de kwaliteit van de wetenschap.



## About the author

Hidde was born on Saturday, 27 April 1991, in Lelystad, the Netherlands. He obtained his bachelor's degree in Social Psychology from Utrecht University in 2013 and graduated from the academic master's program in Social Psychology at the same university in 2014. During his studies, he worked for one and a half years as a student assistant at the Department of Methodology & Statistics at Utrecht University. In this role, he taught various work groups, supervised SPSS computer practicals, and invigilated exams.

After graduation, Hidde held various temporary positions, including roles at the Department of Methods & Statistics as teaching assistant and at Royal HaskoningDHV in the field of traffic and human behavior. In 2015, he was appointed junior lecturer at the Department of Methods & Statistics at Utrecht University. There, he organized exams, assisted in course coordination, and coordinated several courses independently. In addition, he taught work groups, lectures, and practical sessions in various courses within the bachelor's programs in the social sciences. He also taught at University College Utrecht and Marnix Academie.

During this period, Hidde was also briefly involved with other institutions. First, he worked as a consultant for the Workalliance project at the applied sciences university in Utrecht. Second, he was partly employed by Royal HaskoningDHV for four months as a junior advisor. In 2017, he was offered a 50/50 appointment at the Department of Methods & Statistics, combining his work as a junior lecturer with a PhD track.

During Hidde's appointment as a PhD candidate and junior lecturer, he served as a PhD representative on the PhD Council of the Faculty of Social and Behavioral Sciences at Utrecht University. Furthermore, he was involved as a member of the Faculty Network for Qualitative Research from the network's inception. He was also part of the organizing committee of the reading group on teaching statistics.

As of September 2026, Hidde works at the Department of Methodology & Statistics at Utrecht University as an assistant professor with a focus on teaching.



## Publications

### Published manuscripts

**Leplaa, H.J.**, Tönjes, J.A., Bouterse, M., Soppe, K.F.B., & Klugkist, I. (2025). Applying qualitative methods to experimental designs: A tutorial for the behavioral sciences. *PLOS ONE*, 20(6): e0324936. doi: 10.1371/journal.pone.0324936

Heck, D.W., Boehm, U., Böing-Messing, F., Bürkner, P.C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H. A. L., Klugkist, I., Kuiper, R.M., Lee, M.D., Leenders, R., **Leplaa, H.J.**, Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., Mulder, J., ... Hoijtink, H. (2023). A review of applications of the Bayes factor in psychological research. *Psychological methods*, 28(3), 558–579. doi: 10.1037/met0000454

De Ridder, D., Aarts, H., Benjamins, J., Glebbeek, M.-L., **Leplaa, H.J.**, Leseman, P., Potgieter, R., Tummers, L., & Zondervan-Zwijnenburg, M. (2022). “Keep your distance for me”: A field experiment on empathy prompts to promote distancing during the COVID-19 pandemic. *Journal of Community & Applied Social Psychology*, 32(4), 755–766. doi: 10.1002/casp.2593

**Leplaa, H.J.**, Rietbergen, C., & Hoijtink, H. (2020). Bayesian Evaluation of Replication Studies. *Meta-Psychology*, 8. doi: 10.15626/MP.2020.2554

Soppe, K.F.B., Wubbels, T., **Leplaa, H.J.**, Klugkist, I. & Wijngaards-de Meij, L.D.N.V. (2019). Do they match? Prospective students’ experiences with choosing university programmes. *European Journal of Higher Education*, 9, 359–376, doi: 10.1080/21568235.2019.1650088



## Dankwoord

Graag schets ik hier een typische week uit mijn leven. Laten we starten op de maandagochtend, een onderwijsdag, dus ga ik vroeg naar werk. Kelly zorgt dat de kinderen naar opvang en school komen, zodat ik in alle rust vroeg naar werk kan voor de files uit. Daar aangekomen kan ik ongeveer één uur productief werken, dan komt een collega (Anne, Marianne, Thom, Thijs) binnen, en doen we een bakje koffie. Soms sluiten nog wat andere collega's (Gerko, Irma, Kevin, Peter, Rens, Yvonne) kort aan, en we bespreken de belangrijkste bijzaken van het leven.

Dan moet ik onderwijs geven, ik weet dat ik bij terugkomst geïnteresseerde collega's zal treffen (Marieke? Karlijn? Jeltje? Nijs? Rebecca? Collega-JuDio's?) met wie ik kan bespreken hoe het onderwijs verliep. Even fijn ventileren over onze ervaringen, en bespreken hoe de volgende sessie nog aangenamer kan verlopen. Aan het eind van de werkdag spurt ik naar huis, om de kinderen op te halen bij opvang of schoonmoeder. En dan na het eten is er tijd voor work-out via FaceTime, een overblijfsel uit COVID-tijd, met Sameer, Taeke en Dennis.

Dinsdag is mijn vrije dag, sinds mijn ouderschapsverlof in kon gaan heb ik hier aan vast kunnen houden. Samen met Siem of Evie naar het zwembad, koffie drinken, spelletjes doen, en boekjes lezen. Deze dag staat in het teken van hen, met als ze op bed liggen (tegenwoordig: op school zijn) even gelegenheid om wat mails weg te werken of andere korte taken te vervullen. Soms plan ik een vergadering op deze dag, en kan ik rekenen op begrip van mijn collega's als een kind even gedag komt zeggen.

Woensdag is een onderzoeksdag. De kinderen zijn bij mijn vader, die zijn goed onder de pannen. Ik zal vandaag een dag zitten blokken op mijn kantoor en de boterkoek moeten overslaan. Gelukkig is er op kantoor altijd een kamergenoot om even wat stoom af te kunnen blazen als dat nodig is. In de loop der jaren heb ik de nodige kamergenoten versleten: Marieke, Thom, Lientje, Mirjams foto's, Sara, Kirsten, Nikolaj, en nog zoveel meer mensen waarmee ik meer of minder regelmatig het kantoor en onze verhalen mocht delen. Daarnaast zijn er altijd de mensen op de afdeling die achter de schermen alles regelen, het is enorm fijn om te mogen werken met mensen als Els, Marianne, Irma, Chantal, Miro, Flip, Kevin, & Christine.

Woensdag biedt ook ruimte om hard te lopen, met Thom, Thijs, Tobias, Mariska, Lientje, en Peter trekken we er regelmatig op uit in de omgeving van

---

USP. Soms gaan we wat vroeger (dank voor het meegaan Thijs), harder (dank Lientje en Thom), en soms moeten we de veters strikken om verder te gaan (dank Mariska). Maar altijd is het gezellig (dank allen!).

Donderdag verdeel ik mijn aandacht over onderwijs en onderzoek. Ik heb in de ochtend een meeting met een van mijn promotoren. Misschien is de meeting met Charlotte? In het begin van deze reis was ze frequent bij meetings aanwezig, waarbij we samen zochten naar onze rollen. Zij merkte toen al op dat je meer mee kon geven aan mij dan gedacht: en dat heeft ze gedaan. Ik heb een serie one-liners met tips van haar meegekregen, verzameld en ter herinnering op mijn whiteboards gehangen.

Of wellicht is de meeting vandaag met Herbert? Dat belooft een meeting te worden vol met uitwisselingen van visies. Altijd voel ik mij op mijn gemak om mijn gedachten, twijfels en ideeën te spuien. Heb ik een idee waar hij niet aan gedacht heeft, of waar hij niet zelf voor zou gaan? De reactie is: 'Hmmm, dat kan interessant zijn. Kun je uitleggen hoe en waarom dat zou werken?' Positief, constructief, en een open deur om alles uit te spreken.

Maar een meeting met Irene kan ook op de agenda staan. We bespreken dan eerst de persoonlijke zaken: hoe is het met de katten, knieën, en zijn er andere zaken (werk-gerelateerd of niet) om even te delen met elkaar? Daarna hebben we een eerlijke sessie: we bespreken de kritieken, de twijfels, de frustraties. En we bespreken de plannen, de mogelijkheden, de oplossingen.

Halverwege de werkdag is er ruimte om met de collega's te lunchen. Het zijn simpelweg te veel mensen om op te noemen, maar gedenkwaardige momenten waren de Jenga-wedstrijden, discussies over wat een sandwich is, friettafels, gesprekken over wie de mol is, en we kunnen even gevoelens uiten over panenkoeken.

Donderdag is ook een goede avond om, na een borrel (eerst SpAfDr, later wekelijks georganiseerd) uit eten te gaan. Noemenswaardige uitschieters zijn de heerlijke hamburgers met Anne, Marloes, en Marcel, en de bitterballen met Marco & Kathelijn.

Vrijdag is mijn thuiswerkdag. Een dagdeel staat in het teken van het onderwijs van de volgende week voorbereiden. Het andere deel in het zwoegen op onderzoek. Heerlijk om tussendoor even te kunnen hardlopen, met Kelly te kunnen lunchen, en bakjes fruit, wafels, en koppen koffie te krijgen aangereikt van Siem of Evie.

Dat is ook het moment om aan de gezondheid te werken. Specifiek wil ik hier fysiotherapeuten Anton, Maaïke, en Karin noemen, die mij hebben geholpen fysieke klachten te overkomen. Daarnaast psycholoog Marcus, die mij heeft geholpen om mentaal weerbaarder te worden.

In het weekend is er verdere ruimte voor ontspanning. Met de altijd geïnteresseerde (schoon)familie is het een warm gezelschap. Hoewel jullie niet allemaal even thuis zijn in de academische wereld, leidt dat nooit tot verminderde interesse en steun. Altijd staat de koffie (met rozijntjes voor de kids) klaar bij Paul en Andrea, hebben Menno en Jannie tijd tussen de computer- en bridgeclubs voor een koekje, staat er een spelletje klaar bij Monique, Dennis, en Liselotte, of heeft Frank voor Mikey en ons een heerlijk gerecht gemaakt. En om de

---

gedachten te verzetten mag ik jaarlijks mijn twee linkerhanden laten wapperen in Frankrijk: samen met Sameer, Dennis, en Jochem (met na het klussen genoeg proeverijtjes en barbecues).

Al de mensen hierboven genoemd, en de vele mensen die niet bij naam zijn genoemd maar ook in dit verhaal passen, hebben een bijdrage geleverd aan dit proefschrift. Door hen is de werkomgeving zo prettig, is er steun vanuit allerlei kanten, en is er altijd een geïnteresseerd en luisterend oor te vinden. Ik verkeer in een enorme luxe-positie met een dergelijk aangename werkomgeving, en zulke lieve mensen in mijn persoonlijke leven.

In het bijzonder wil ik hier nog een aantal mensen bedanken. Mijn paranimen, Thom en Anne, zijn twee van de collega's die als een rode draad door mijn PhD-carrière liepen. Anne, we zijn vlak na elkaar gestart in deze rol. Ik heb altijd genoten van onze gesprekken aan de start van de werkdag. Het is enorm bijzonder hoe wij ook nu nog, jaren nadat je bij M&S bent vertrokken, elkaar nog zo goed als wekelijks treffen. Thom, in de tijd van COVID en (gedeeltelijke) lockdowns was jij een van de weinigen met wie ik nog wel menselijk contact had. In die laatste jaren was het zo fijn om tegen je te mogen klagen, om elkaar te helpen alles te relativiseren, of gewoon te babbelen over alles in het leven. Ik ben enorm blij met jullie beide: jullie zijn een groot deel van de fijne werkomgeving voor mij.

Yme & Zoë, samen met Lev, Mats, en Fay, dagen mij uit om anders naar de wereld te kijken, en ze houden mij nog fit ook. Merle & Michael, met natuurlijk Foss en ....., tonen mij het belang van altijd klaar staan voor een ander, en brengen daarnaast veel gezelligheid.

Daarnaast natuurlijk specifiek aandacht voor mijn ouders, op wiens steun ik altijd kan rekenen. Of ik besluit om de ene of de andere kant op te gaan, ik weet dat zij naast, achter, en voor mij zullen staan. Een warmer nest is niet te vinden.

Siem en Evie, die mij hebben geleerd om alles in perspectief te zien. Mooi dat prioriteiten zo makkelijk kunnen verschuiven, en wonderlijk hoeveel liefde één hart aan kan.

Kelly, die zoveel geduld met mij heeft gehad, en dat nog steeds heeft. Zorgen wegnemen, taken op je nemen, en af en toe een vriendelijke schop onder mijn kont. Wat ben ik ongelofelijk blij met jou aan mijn zij.

Maar het laatste woord wil ik besteden aan Herbert, Irene, en Charlotte. Waar de hiervoor genoemde mensen ervoor gezorgd hebben dat ik mentaal (redelijk) gezond ben gebleven, hebben jullie een essentiële rol gehad in dit proefschrift. Naast de aandacht voor mijn mentale gezondheid die jullie hadden, hebben jullie altijd meegedacht én proactief suggesties gedaan om te kijken hoe we dit proces tot een goed einde konden brengen. Ik durf te zeggen dat ik zonder jullie *wel* gelukkig was geworden, daar zorgen de andere mensen in mijn leven ook voor, maar dat ik *niet* op deze plek in de academische wereld terecht was gekomen. Dank aan alle drie voor jullie rol in dit proces!



---

*‘When you don’t have a hammer,  
you don’t want anything to look like a nail.’*  
Kagan, in *Of Paradise and Power* (2003)

