ALJAŽ SLUGA

Hour of Judgment

On judgment, decision making, and problem solving under accountability

Hour of Judgment

On judgment, decision making, and problem solving under accountability

Hour of Judgment
On judgment, decision making, and problem solving under accountability

Uur der waarheid
Over oordelen, besluitvorming en probleemoplossing onder accountability

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof. dr. F. A. van der Duijn Schouten

and in accordance with the decision of the Doctorate Board.

The public defense shall be held on

Thursday 1 July 2021 at 13:00 hrs

by

Aljaž Sluga
born in Postojna, Slovenia

Erasmus University Rotterdam

Doctoral committee

Promotor:
Prof. dr. F. G. H. Hartmann

Other members:
Prof. dr. E. Peek
Prof. dr. G. Jacobs
Dr. B. de Langhe

Copromotor:
Dr. M. A. S. Boksem

Erasmus Research Institute of Management - ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam
Internet: https://www.erim.eur.nl

ERIM Electronic Series Portal: https://repub.eur.nl/pub

ERIM PhD Series in Research in Management, 520
ERIM reference number: EPS-2021-520-F&A
ISBN 978-90-5892-605-0

Cover art: ©Julia Olaussen
Cover design: PanArt, www.panart.nl

This publication (cover and interior) is printed by Tuijtel on recycled paper, BalanceSilk®
The ink used is produced from renewable resources and alcohol free fountain solution.

Certifications for the paper and the printing production process: Recycle, EU Ecolabel, FSC®, C007225
More info: www.tuijtel.com

Acknowledgments

This text completes the project that brought me to the wonderful city of Rotterdam in August 2015. Roughly half a decade later it's complete, or at least as complete as it will ever be. I'd like to express my gratitude to everyone who's helped me along the way. Apart from Mr. Cohen—who kindly refrained from objecting through his little window above as I planted his quotes in my chapters—the list includes at least the following people and a cat.

My supervisors, Frank and Maarten, were always willing to dedicate time and energy to our project. The conversations we had throughout the trajectory were never boring and taught me a great deal. Occasionally we succeeded in staying on topic, but replicating those experiments proved to be difficult. I'd like to thank Frank and Maarten for their thoughtful guidance, persistent encouragement, and doing their best to get the balance of accountability right.

I would also like to thank the members of the Assessment Committee for reviewing the dissertation. My gratitude extends to everyone who commented on the drafts that preceded it—these and other chapter-specific acknowledgments are included at the start of each chapter.

I owe a lot to Iuliana, who kindly tolerated my continued coffee-drinking presence in her office. Paul Erdős once abstained from coffee for a month and afterward concluded that the experiment set mathematics research back for the entire period. My *not* abstaining from coffee may have indirectly done something like that to research in accounting education. Among other things, I cherish the memories of our shared discovery of all things Bayesian. Then Iuliana had a baby and that was that. Little Remy will have to carry this burden all his life.

I thank every other colleague in the department, including my PhD colleagues Marcel, David, Anke, Sebastian, Karin, Albert, and Anoek. I will remember my time among the accountants (and Sandra) fondly. I hypothesize that the sound of laughter of some of these colleagues could be used to pierce tunnels through the Alps and I live in fear that one day someone will weaponize it. My thanks extend to other colleagues at the school, especially the Erasmus Behavioural Lab crew—Christiaan, Marcel, Gerrit-Jan, and Mischa were very helpful on more than a few occasions.

I need to re-mention Albert, without a question my favorite Hungarian insect and a worthy adversary in *calcettino*. Albert's game nights supplied many a fun memory over the years and I've greatly enjoyed our office chats and trips to Cinerama, where a small genocide may have been committed on their popcorn population. I have always appreciated Albert's active disposition and even his embarrassingly unsuccessful attempts to get me to do sports that require a larger range of motion than calcettino.

Then there is the mercurial Philip, who I could always count on for a delightful conversation, dubious wisdom, and a letter from Greece that never came. I cherish the memories of our grand chess match that spanned six years and four cities. What could be less important than the fact that the final score was 15.5:5.5 in my favor? I'm certainly not the sort of chap who'd dig through email, photos, chat history, pdf documents, *and* text notes to uncover and broadcast the fact that this chess match ended 15.5:5.5 in my favor. But I'm grateful to Philip for many things, not just partaking in a chess match that ended 15.5:5.5 in my favor. He's also the reason I know Julia, the enigmatic Norwegian artist who painted the cover of this dissertation. I'd like to thank Julia for her work & patience and urge the reader to commission her when the need arises—she also does great nudes.

On the other side of the Alps, there's my family, who made sure my head was not completely turned by the two majestic triumphs of Dutch culture, namely *hagelslag* and *poffertjes*, and who virtuously resisted the urge to turn my room into a hunter's trophy room—it would have been difficult to sleep next to that taxidermied baby bear. For that and much more, I thank them. Frankly, I'm not so sure I want to extend my thanks to our cat Čička, whose statistical advice was often subpar. But I will anyway for her uncomplicated companionship and the purring so sweet that it would heal this world if only the world would listen.[1]

Finally—and most of all—Sara. Her love is my safe harbor and her steadfast conditioning has improved me significantly over the years. My behavioral repertoire is now much richer, full of exciting new operants. For example, gone are the days when I'd just place a cup of tea directly on the table like some Viking. I'm now intimately familiar with coasters and use them with joy and precision. Sara has been an ever-present source of support and a brilliant coauthor to our little Dutch adventure. I cherish the memories of our trips; the small ones inside the kingdom, the bigger ones outside its borders. Most often, though, I catch my mind revisiting the tiny ones; the leisurely Saturday strolls revealing mysterious bits of Rotterdam hiding within walking distance of our couch, the relentless hunts for milkshakes. Big, small, tiny—Sara made every trip more special. She's also someone who stoically read the entire draft of this dissertation (yes, even the footnotes) and put some limits on the grammatical creativity lurking within the pages. For all of these things and much more, I thank her. But especially for the coaster thing.

[1] It's also louder than a tractor.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

I know that there is an eye that watches all of us
There is a judgement that weighs everything we do

— Leonard Cohen knows about annual performance reviews

Summary:

This chapter provides the context and introduction to the rest of the dissertation. After introducing the topic of this dissertation and delimiting its scope (section 1.1), I consider definitions of accountability and settle on two working definitions, a broad one and a narrow one (section 1.2). Next, I summarize the previous experimental literature on the topic and introduce and discuss the distinction between process and outcome accountability (sections 1.3 and 1.4). Sections 1.4, 1.5, and 1.6 highlight the issues and questions that motivate the empirical research reported in Chapters 2, 3, and 4. In particular, section 1.4 emphasizes the importance of replication and introduces Chapter 2. Section 1.5 notes the paucity of research on the effects of accountability on creative problem solving, introduces insight problem solving as one aspect of creative problem solving, and outlines Chapter 3. Section 1.6 discusses the connection between (ir)rationality and accountability and motivates Chapters 4 and 5. Next, section 1.7 discusses how the value of laboratory research can be maximized and concludes with five recommendations for future research. Finally, section 1.8 consists of a declaration of contribution.

Acknowledgments:

I thank Philip Eskenazi for commenting on an earlier version of this chapter.

1.1 Scope of the dissertation

The studies in this dissertation deal with human judgment, decision making, and problem solving. Each of these topics is incredibly broad and has spawned a correspondingly vast literature in psychology, management, neuroscience, and other academic fields. But much of this work ignores the social context in which these activities of the human mind take place. As Tetlock observed back in 1985, laboratory studies tend to create a "social vacuum" in which experimental subjects have no need to consider the interpersonal consequences of their actions. Outside of the laboratory, in contrast, people frequently make decisions that are scrutinized by others, report on and defend their approach to problem solving, and justify their judgments in order to be evaluated favorably by others and not face unpleasant consequences. In short, people are *accountable* to each other. Research conducted since 1985 testifies to the importance of understanding this social context when making predictions about human judgment, decision making, and problem solving (Lerner and Tetlock, 1999; Hall et al., 2017). This dissertation presents an attempt to contribute to this body of work.

I approach the topic of accountability from the standpoint of a management system designer who is interested in facilitating the performance of individuals. I use the term *management system designer* in a broad sense; it subsumes managers interested in influencing the behavior of their direct reports, executives concerned with the behavior of a great many members of their organization, and even supraindividual entities such as oversight boards. And by *performance* I mean any kind of desirable external outcome that is at least partly under the control of the accountable individual, including successful hiring and other business decisions, useful and accurate forecasts, fair performance appraisal ratings, creative marketing campaigns, etc.

For the management system designer, various ways of introducing accountability are part of a larger toolkit that includes approaches such as educational campaigns and training (for example, training managers to use *natural frequencies* to improve probabilistic reasoning; Hoffrage et al., 2015), changes to the physical work environment (for example, switching to or ditching an open office plan; Bernstein and Waber, 2019), setting default options and various other forms of *choice architecture* (Thaler and Sunstein, 2009), etc. In order to make an informed decision about whether to rely on accountability instead of (or together with) these other approaches, the management system designer needs to know under what circumstances, and in what form, accountability facilitates performance.

Questions like these can be difficult to answer for multiple reasons. For example, it can be difficult to disentangle the effects of a particular accountability system from other causal forces. And when other individuals or groups of people carry out the proposed system, monitoring the quality of the implementation can be challenging. Outcomes of interest can be difficult to measure, especially as we move away from simple types of productivity and wish to influence more complex outcomes, such as creative output. Outcomes of interest can also take a long time to manifest, limiting the scope for exploration and raising the stakes of introducing a change to the status quo.

Laboratory experiments present one way of cutting through this complexity and obtaining answers to difficult questions. Bringing accountability into the lab has some compelling advantages. We can make justified inferences about causality. We can manipulate accountability

in (almost) any way we like, even if it is not something real-world organizations are particularly keen on trying. We can administer detailed measures to gain insight into the underlying mechanisms that drive behavior. And we can repeat our experiments rather easily to see if the results hold up. Clearly, laboratory experiments are not the only way to study accountability questions, neither do I claim they are the best. But they do present a kind of a playground where ideas on how accountability should be implemented can be investigated systematically and at a relatively low cost. From this vantage point, the observation that accountability in the lab often differs a lot from accountability in the real world can be reframed from a criticism to an expression of approval. Less of a "major methodological limitation" (Patil et al., 2014), more of a potential step toward something better.[1]

The approach I follow in this dissertation (both content-wise and method-wise) means that many things are left out. For example, the focus on the *individual* means that the performance of teams under accountability is ignored, even if team-level accountability sometimes works better than individual-level accountability (Garbers and Konradt, 2014). And the focus on *performance* means that the effects of accountability on the well-being of accountable individuals are likewise ignored, even if they are considerable, a precursor to future performance, and important in their own right (Hall et al., 2006; Hochwarter et al., 2007). These and many other related topics are important but they are not covered in this dissertation, except indirectly.

1.2 In search of accountability

The larger academic literature offers numerous definitions of accountability, leading Mulgan (2000) to describe it as a "complex and chameleon-like term" and an "ever-expanding concept." He notes several ways in which scholars in the field of public administration extended the concept beyond its original, core meaning. These extensions include *internal* accountability to oneself in the absence of external scrutiny, accountability as a catchall term for all forms of *control* that one party uses to bring the behavior of another party in line with its interests, accountability as the degree to which one party is *responsive* to the preferences of another, and accountability as *dialogue* that consists of open discussion and debate (Ibid.). In academic fields closer to this dissertation, the desire for accountability to engulf ever larger regions of conceptual space is less pronounced. Nevertheless, some variation exists, making it important to clarify what the term captures in this dissertation.

In line with Mulgan (2000), I highlight the potential for *external scrutiny* as a defining component of accountability. The accountable individual believes that an external party may monitor and evaluate their behavior or its traces and consequences. Moreover, the accountable individual believes that non-negligible consequences for their well-being may stem from this evaluation. Therefore, the external party (*evaluator*, *principal*, *audience*—I use these terms interchangeably in this dissertation) can be functionally decomposed into two elements. First, the evaluator *assesses* the acceptability of the accountable individual's behavior or its consequences and in this way serves as a mediating variable between the individual's output and the

[1] This argument should not be abused. It only pertains to those pecularities of accountability design found in the lab that could potentially be exported to real-world accountability systems.

rewarding or punishing consequences for the accountable individual. In some accountability systems, the outcome of this assessment may eventually be formalized, as when a supervisor assigns a numerical rating to the performance of their direct report in a given year. Second, the evaluator is a source of rewarding or punishing *consequences*. These consequences could be tangible, as when a supervisor decides whether their direct report should be demoted (or promoted). The consequences need not be immediately tangible, however (Siegel-Jacobs and Yates, 1996). Given the importance of other people for well-being or even survival, individuals are highly attuned to the reputational consequences of their own behavior (we can think of these as deferred tangible consequences). In essence, then, accountability exploits our ability to anticipate (through *mentalizing* and associated processes; Frith and Frith, 2006) these two components of accountability and adjust our thought and behavior in advance.

One element that is conspicuously missing in the previous paragraph is *account-giving*, the process of describing, explaining, and justifying one's behavior or its consequences. This element is central to the original meaning of accountability (Mulgan, 2000) and many of its definitions. For example, the most often quoted definition in the psychology literature defines accountability as an "implicit or explicit expectation that one may be called on to justify one's beliefs, feelings, and actions to others" (Lerner and Tetlock, 1999). On the other hand, some definitions leave account-giving out. For example, Hall and Ferris (2011) define it as "an implicit or explicit expectation that one's decisions or actions will be subject to evaluation by some salient audience(s) with the belief that there exists the potential for one to receive either rewards or sanctions based on this expected evaluation." Or, as Merchant and Otley (2006) succintly put it, accountability means that "individuals are rewarded when good things happen and punished when bad things happen." Moreover, in the experimental literature accountability is sometimes operationalized in a way that does not include account-giving (see section *Process and outcome accountability*).

In light of this variation, I offer a *narrow* conception of accountability that includes account-giving as a defining component and a *broad* conception where account-giving is optional. From an *ecological* perspective, spilling ink over this distinction may seem unneccessary because the potential for account-giving seems almost unavoidable. But from a *scientific* perspective, we should not prematurely decide it is irrelevant. For example, if some experimental manipulations feature account-giving and others don't, we need to recognize the difference and the possibility that it matters. And from a *design* perspective, the presence, form, and intensity of account-giving is one more set of knobs we can turn when seeking to incentivize optimal performance.

From the perspective of this dissertation, accountability is a social technology, a broad spectrum of tools available to the management system designer characterized by their common purpose and means of achievement. Many investigators, however, use the term *felt accountability*, which draws attention to the psychology of the accountable individual. The term serves a useful purpose in reminding the management system designer that the effects of a formal accountability system ultimately hinge on the interpretations and characteristics of the individuals interacting with it (Hall et al., 2017). It need not, however, involve any kind of ontological commitment as to what kind of a psychological state accountability is or even that such a distinct state exists.

A more ambitious usage of the term is found in survey research, where felt accountability is a

construct often measured with the scale developed by Hochwarter et al. (2003, see Hochwarter et al. (2007); Hall et al. (2017)). The scale features eight items that refer to intensity (e.g. "The jobs of many people at work depend on my success or failures") and scope (e.g. "I often have to explain why I do certain things at work") of accountability arrangements being present in the individual work environment. Such instruments are used to correlate felt accountability with other variables of interest, such as job performance (Vivian Chen et al., 2016), job satisfaction (Breaux et al., 2009), job-related tension (Hochwarter et al., 2005), and abusive behavior (Mackey et al., 2018). These self-report measures of felt accountability are relatively blunt research tools, designed to fit the needs and characteristics of survey research. In contrast, the high degree of control afforded by the laboratory approach means there is less need to capture individuals' subjective impressions. Accordingly, there is little measurement of felt accountability in this stream of research and in this dissertation (but see Chapter 6).

1.3 A short tour of the literature

Accountability has been the subject of laboratory experimentation for over four decades. What has been learned so far that could assist the management system designer in their quest to improve the performance of individuals? In this section, I summarize the main theoretical ideas in the literature while also highlighting a few experimental demonstrations and relating these ideas to the perspective of a management system designer that underlies this dissertation. For the purpose of this review, I divide the literature into two periods. *Early period* spans from the initial research conducted in the seventies up to the publication of the influential review by Lerner and Tetlock (1999), and *later period* refers to research conducted ever since.

The results of the early period are encapsulated in the *social contingency model of accountability* advanced by Tetlock (1985; 1992) and later updated by Tetlock and Lerner (1999). The model is rooted in an inductive approach to theorizing and remains the most influential theoretical account of accountability in the psychology literature (Hall et al., 2017). As its name suggests, the model emphasizes the varying, contingent nature of the effects of accountability. Because these effects can run counter to the interests of the management system designer and other relevant stakeholders, the model underscores the need to carefully consider the decision to implement accountability and the way in which this is accomplished.

Before summarizing these varying effects and their precursors, let us briefly examine the core of the social contingency model. First, the model views behavior as being driven by explicit motives. Although the model inherits five basic motives from the broader literature, it mainly emphasizes two. These are *seeking social approval* and *minimizing cognitive work*.[2] Accordingly, individuals are assumed to think and act in ways that are likely to result in the approval from the relevant audience while also minimizing the extent of cognitive work. Second, the model's implicit area of application is the accountable individual's thought process and behavior during a discrete episode composed of three stages: *(i)* initial cognitive activity and behavior that leads to *(ii)* some externally relevant output (such as judgment or choice) and is followed

[2] The ones not mentioned in the main text are achieving cognitive mastery, maximizing the cost-benefit ratio of relationships, and asserting autonomy.

by *(iii)* further cognitive activity that cannot change the output anymore. As an example, a loan officer could *(i)* assess a loan application, *(ii)* make the call to approve a loan, and *(iii)* continue receiving relevant information afterwards. Against this background, the model emphasizes how characteristics of accountability (as reflected in the accountable individual's *state of knowledge*) and other relevant factors lead to different *coping strategies* (patterns of thought and behavior) by moderating the relative importance of different motives and the feasibility of their achievement (Tetlock and Lerner, 1999).

The initial formulations of the model highlighted three major coping strategies and linked them to variation in two aspects of an individual's state of knowledge regarding the accountability situation they find themselves in. These aspects are the *timing* of the accountability imposition and the extent of the accountable individual's knowledge about the *audience's views*. Starting with the latter, the model suggests caution when introducing accountability to an audience with a directional preference that does not coincide with the objectives of the management system designer. In this case (as long as the imposition of accountability happened before stage *(iii)*), the model suggests that individuals are likely to engage in *(strategic) attitude shifting* in the direction of the audience's view or preference, as this coping strategy jointly maximizes social approval and minimizes cognitive work (on the assumption that the audience is less likely to demand rigor if the output agrees with their preference). A classic case appears in the performance appraisal context (Klimoski and Inks, 1990). The management system designer may wish to facilitate ratings that accurately capture employees' performance but the rating manager's accountability towards their direct reports (who have a clear preference for higher ratings) often results in elevated ratings (*leniency bias*; Bretz et al., 1992; Bol, 2011).

As an antidote to this tendency towards conformity (that may not neccessarily be strategic; Pennington and Schlenker, 1999), the model highlights accountability to an audience with *unknown* views (a condition of so-called *normative ambiguity*; Tetlock, 1992). The resulting coping strategy (*pre-emptive self-criticism*) involves effortful, multi-dimensional, and self-critical cognitive processing (Tetlock, 1985). An early demonstration was provided by Tetlock (1983a) who asked his experimental subjects to take a position on controversial social issues. While subjects' responses tended in the direction of the audience's views when the latter were known, private thought protocols revealed more *integratively complex* thinking when the views of the audience were unknown. The benefit of this kind of accountability also interacts with timing; Tetlock (1983b) found it reduced the *primacy effect* (undue influence of information presented early on the final judgment) but only when subjects were made aware of the need to justify their judgment before they began evaluating the evidence, i.e. before stage *(i)*. This kind of accountability, Tetlock (1983b) concludes, "may represent a simple, but surprisingly effective, social check on many judgmental shortcomings documented in the current experimental literature." Returning to our example of performance appraisal, knowledge of the audience's preference seems difficult to avoid in this setting (although it may be exacerbated with the practice of *ex-ante* self-assessment) as does a manager's accountability towards their direct report. Following the social contingency model, one solution would be to introduce additional accountability to an audience with incongruent preferences. The practice of *calibration committees* presents one way of accomplishing this (Demeré et al., 2018).

The discussion so far focused on so-called *predecisional* accountability. That is, individuals are aware of the need to justify their position or decision before they commit to it. If accountability

is instead imposed retroactively (stage *(iii)* above), it is likely to facilitate efforts to justify the previous position or decision (*defensive bolstering;* Tetlock et al., 1989). Accordingly, accountability can exacerbate commitment to a failing course of action (Simonson and Staw, 1992)[3]. The business simulation study by Fox and Staw (1979) suggests that this effect is strongest when the individual is accountable to a skeptical audience and has low job security.

Overall, the social contingency model has clear practical implications. It sensitizes the management system designer to the danger of retroactive accountability and the importance of considering the preferences of relevant audiences. The latter aspect also makes it clear that decreasing transparency and information flow can be a valuable addition to the toolkit of the management system designer. At the same time, the model's implications seem unlikely to be particularly revelatory to an experienced practitioner, limiting its utility as a roadmap to management system design.

Moving onto the later period of experimental research on accountability, we can detect two main streams. One of these is primarily concerned with phenomena in which human cognition appears to fall short of some normative standard. This work investigates if accountability can be used to eliminate or at least attenuate a particular bias or error (typically as part of a broader research agenda concerning said bias or error). Because earlier literature (summarized above) highlights the potential of what Lerner and Tetlock (1999) describe as *predecisional* (and *preexposure*) accountability to an audience with *unknown* views, this stream employs this particular form of accountability. For example, researchers have investigated if accountability reduces judgmental biases in the context of negotiation (de Dreu et al., 2000), decreases predecisional distortion of information (a form of confirmation bias) by auditors and salespeople (Russo et al., 2000), curtails self-enhancement tendencies (Sedikides et al., 2002), decreases loss aversion (Vieider, 2009), reduces preferences reversals induced by framing (Vieider, 2011), decreases escalation of commitment (Moser et al., 2013), improves information sharing and decision quality in a group decision-making context (Scholten et al., 2007), and improves decision making involving simple and compound events (Vieider, 2011). In the field of management accounting, Libby et al. (2004) examined if accountability improves performance evaluation judgments based on balanced scorecards, Rausch and Brauneis (2015) investigated the potential of accountability to facilitate optimal preparation of information for managerial decision making by management accountants, and Fehrenbacher et al. (2020) tested the ability of accountability to reduce the influence of affective reactions to a manager proposing a project in capital budgeting decisions. Overall, most of these studies highlight the potential of accountability as a tool for improving performance.

The second dominant stream of recent experimental literature focuses on the distinction between *process* accountability and *outcome* accountability. Because the distinction plays a prominent role in this dissertation, it is covered separately in the next section.

[3] But see Schulz-Hardt et al. (2020) for a recently published failed attempt to replicate this effect.

1.4 Process and outcome accountability

The body of work summarized in the previous section suggests that accountability can be an effective tool for debiasing faulty judgment or choice and otherwise improving performance. In line with the social contingency model, these effects are typically observed when researchers implement predecisional, preexposure accountability under the conditions that prevent simple conformity to a flawed standard. But the literature also hints at another characteristic of successful accountability systems. In particular, researchers often explicitly inform their experimental subjects that they will have to explain and justify *how* they arrived at a particular judgment or choice. Under this type of accountability, the evaluative focus is on the process(es) that an individual employs during judgment, decision making, or problem solving. Accordingly, this type of accountability is called *process* (or *procedural*) accountability (Simonson and Staw, 1992; Siegel-Jacobs and Yates, 1996; Patil et al., 2014).

To the management system designer, however, the process that the accountable individual relies on is not in itself of primary importance. Naturally, there will always be constraints (for example, ethical ones) that the accountable individual's way of doing things should conform to. As long as this condition is satisfied, however, what the management system designer ultimately cares about is performance, or the various external *outcomes* that are desired by the relevant stakeholders. This observation immediately suggests the idea of *outcome accountability*—a form of accountability in which the evaluative focus is on the outcomes rather than the processes used to achieve these outcomes (Ibid.). This distinction is relevant when external outcomes are usefully expressed as a function of two kinds of factors, with *controllability* as the dividing line. *Process* concerns those activities, physical or mental (i.e. conspicuous or not), that are under the individual's control. The remaining contribution is due to "external forces" that are *not* under the individual's control.[4]

In the experimental literature, process accountability is operationalized by informing subjects in advance that they will be asked to describe, explain, or justify their process (reasoning, strategy) in writing or in an interview-like setting. In contrast, outcome accountability is operationalized in two ways that implicitly draw on the different conceptualizations of accountability (*narrow* and *broad*) articulated in the previous section. In the first version of outcome accountability (that satisfies the narrow definition of accountability as well as the broad one), account-giving is present and focused on the outcome(s). For example, outcome-accountable subjects in the negotiation study of Chang et al. (2013) were required to write a report that would justify their negotiation outcome (the report "needed to address both the negotiation result and whether the outcome was considered acceptable for their company"). In the second version of accountability (that satisfies the broad definition but not the narrow one), this account-giving component is absent and the basis for external evaluation is purely the realized outcomes. To take an example from another management accounting study, outcome-accountable subjects

[4] This bright line of demarcation can be pale and murky in practice. A business can, for example, take steps to maximize the robustness of its operation to otherwise uncontrollable external threats such as a global pandemic. To take a more subtle example, consider a pernicious cognitive bias that is shown to be impervious to intent and effort. The accountable indvidual could still learn about it and take steps to bypass it, for example by using a decision aid or controlling the information environment. This example shows why the phrase "external forces" needs the scare quotes.

of Dalla Via et al. (2018) making an investment decision for their hypothetical company were informed that their "evaluation as a manager only depends on how close your decision is to the most optimal decision."

For the management system designer, outcome accountability seems to potentially offer some compelling advantages. One of these relates to the fact that outcomes are of direct interest to the designer whereas processes are not. The immediate *evaluative focus* of the accountability system plays a powerful incentivizing role in directing the behavior of the accountable individual. Individuals may, intentionally or not, prioritize activities that are reflected in this focus. Accordingly, the lesser the correspondence between the ultimate objectives of the management system designer and the focus of the accountability system, the more likely it is that the accountability system will not serve its purpose well. This issue appears more likely to arise under process accountability. Another potential advantage of outcome accountability is that it may be more likely than process accountability to facilitate out-of-the-box thinking, experimentation, and creativity in finding ways to tame seemingly uncontrollable outcomes (Patil et al., 2014). Finally, outcome accountability systems may be more transparent and have fewer implementation costs, as the desirability of the outcome is often easier to assess than the quality of the process that was used.[5]

These potential advantages of outcome accountability notwithstanding, experimental research gives reason for caution as direct comparisons tend to favor process accountability. An early demonstration was provided by Simonson and Staw (1992) who found that process-accountable individuals allocated fewer further resources to a failing course of action than outcome-accountable ones[6] and suggested that process accountability may serve as a "potent deescalation technique." Other early results are reported in Siegel-Jacobs and Yates (1996) who found more accurate probability judgment under process accountability than under outcome accountability, and Doney and Armstrong (1996) who, in a rare correlational study, found that self-reported process accountability among organizational buyers correlated with higher efforts to gather and analyze the relevant information whereas the same wasn't true for self-reported outcome accountability. These early comparisons were later extended by a growing chorus of studies that looked into counterfactual thinking and affective reactions to decision outcomes (Markman and Tetlock, 2000), judgment of leadership potential (Brtek and Motowidlo, 2002), the *decoy effect* in decision making (Slaughter et al., 2006), multiple-cue learning (de Langhe et al., 2011; Patil et al., 2017), customer-supplier negotiations (Chang et al., 2013), self-serving behavior (Pitesa and Thau, 2013), professional skepticism during an audit (Kim and Trotman, 2015), idea generation (Häusser et al., 2017) forecasting of real-world events (Chang et al., 2017), project funding decisions (Dalla Via et al., 2018), and the trade-off between exploration and exploitation (Verwaeren, 2019). With some exceptions, the results of these studies corroborate the early impression that holding individuals process accountable is preferable to the more direct approach of holding them outcome accountable. An important qualification is that if process accountability is paired with a specific normative standard *and* that standard is

[5] I have added the qualifier "potentially" at the beginning of the paragraph because each of the stated advantages of outcome accountability that follow smuggles in a host of assumptions. Consider, for example, that the outcomes that the management system designer wants to facilitate may be difficult to assess. If the designer then introduces a poor-enough proxy as the basis of an outcome accountability system, the potential of the latter for distortion and *surrogation* may be as high or even higher than that of process accountability alternatives (Choi et al., 2012).

[6] But see Schulz-Hardt et al. (2020) for a recently published failed attempt to replicate this effect.

inappropriate for the task at hand, outcome accountability results in better performance (Patil et al., 2017).

These results have started to influence recommendations for practice. For example, Peecher et al. (2013) draw in part on this body of work to set up a 2 (rewards versus penalties) × 2 (processes versus outcomes) accountability framework for the domain of financial statement audit aimed at regulators interested in facilitating audit quality. Similarly, Schillemans (2016) draws on these and other findings from the experimental literature on accountability to provide recommendations for structuring accountability in the public sector.

But although the number of experimental comparisons of process and outcome accountability has grown steadily in the last two decades, many questions remain open. The *where*, *how*, and *what* of the performance effects of process and outcome accountability are still insufficiently explored. As Hall et al. (2017) emphasize, our understanding of where (for example, in which tasks and under which circumstances) one type of accountability will prove more effective is limited and future research therefore needs to "explore the boundary conditions under which process or outcome accountability produces better outcomes for the accountable actor and the organization."

The limited understanding of the *where* points toward a limited understanding of the *how*. The conclusion that process accountability typically leads to better performance than outcome accountability is more of an inductive generalization of experimental results than a claim rooted in clear understanding of the underlying psychological mechanisms. Several candidate explanations have been offered, however. Siegel-Jacobs and Yates (1996) suggested that *controllability* is lower under outcome accountability and that the corresponding greater chance of failure results in *stress* which then deteriorates performance (Janis and Mann, 1977). But while accountability undoubtedly can result in stress (Hall et al., 2006) and stress undoubtedly can affect performance (Starcke and Brand, 2012), there is no evidence that stress mediated any of the reported experimental findings in this literature. To the best of my knowledge, only one study measured stress and the results revealed comparable levels of stress under process and outcome accountability (Häusser et al., 2017). Another suggestion of Siegel-Jacobs and Yates (1996) was that process accountability provides *guidance* whereas outcome accountability does not. But it is precisely the open-ended, normatively ambiguous form of process accountability that has been investigated the most and that tends to result in better performance than outcome accountability (Patil et al., 2014). It seems implausible that the minimal degree of guidance such a manipulation provides would prove to be consequential (Lerner and Tetlock, 1999). More recently, Patil et al. (2014) proposed that variation in *relational micro-signals* (for example, if the accountability manipulation is phrased in terms of a demand or a polite request) can lead to "major effects on the cognitive, emotional, or behavioral responses of agents to accountability demands," deeemed it "quite plausible that previous studies have conveyed more positive relational signals to those under process accountability than to those under outcome accountability", and suggested that the resulting effects on *psychological empowerment* may mediate the observed effects on performance. Finally, several researchers attributed the diverging effects of process and outcome accountability to the *amount of information* considered by the accountable individual or the *nature of cognitive processing* operating on this information (Siegel-Jacobs and Yates, 1996; Brtek and Motowidlo, 2002; Chang et al., 2013; de Langhe et al., 2011; Dalla Via et al., 2018). The differences concerning the latter can be

couched in the language of dual-process theories which distinguish between processes of *Type 1* and *Type 2* (Evans and Stanovich, 2013). The former are described as implicit, automatic, and unconscious, whereas the latter are described as conscious, deliberate, effortful, and mediated by working memory (Kahneman, 2011). Within this framework, process accountability would be associated with greater reliance on the processes of Type 2 (de Langhe et al., 2011). This class of explanations has received the most support (see Chapter 4), although some evidence is mixed. For example, Dalla Via et al. (2018) used eye-tracking to find greater search effort under process than under outcome accountability whereas Siegel-Jacobs and Yates (1996) reported no differences in information acquisition between the two conditions using a related process-tracing technique. Overall, the scope and precise nature of these information processing effects remain in need of clarification.

Finally, the *what* in the "*where*, *how*, and *what* of the performance effects" refers to the features of accountability manipulations that contribute the most to a given finding. Patil et al. (2014) suggest that "the effects of process and outcome accountability hinge less on the process–outcome distinction than they do on the social-identity signals that accountability sends to employees about how managers view them." Other, more easily observable features of accountability manipulations that sometimes differ (and sometimes not) between operationalizations of process and outcome accountability include the presence of account-giving and the presence of a financial incentive. It is therefore important to refrain from automatically assuming the equivalency of manipulations that in one way or another emphasize the process (outcome). Another implication of this issue is the need to carefully record and share the full experimental materials used in a study. Widespread adoption of this practice would facilitate the generation and testing of hypotheses such as that of Patil et al. (2014). A researcher could, for example, collect the experimental materials, rely on an independent sample of subjects to measure how the materials are perceived, and examine if a measure of a construct of interest, such as controllability (Siegel-Jacobs and Yates, 1996) or empowerment (Patil et al., 2014), predicts the reported effect sizes.

Three of the chapters in this dissertation compare the effects of process and outcome accountability on cognition and performance. In Chapter 2 we note the almost complete lack of independent replication studies in the literature on process and outcome accountability and make the case for the importance of replication attempts in accountability research. We then report the results of a study that compared the accuracy of multiple-cue judgment under process and outcome accountability, replicating and extending several key findings from the literature. In particular, we investigate if process-accountable subjects develop more accurate judgment in multiple-cue tasks that call for *cue abstraction* (de Langhe et al., 2011), if this effect is mediated by so-called *rational thinking style* (Novak and Hoffman, 2009), and if the effect holds when some of the cues are not informative. In addition, we analyze additional measures of decision time and explicit task knowledge and explore if the second-order probability judgments of process-accountable individuals are more accurate and better calibrated than those of outcome-accountable individuals (Siegel-Jacobs and Yates, 1996). Jointly, the results of these analyses provide information on the replicability and generalizability of the relative effects of process and outcome accountability. The other two chapters on the topic investigate the effects of process and outcome accountability on insight problem solving and adaptive decision making. These are introduced in the next two sections.

1.5 Creativity and accountability

Experimental studies on accountability tend to focus on phenomena of judgment and choice, typically asking their subjects to select a response from a predetermined set. This holds for older as well as more recent studies, and for studies comparing accountability with its absence as well as studies comparing the different ways of arranging accountability. What underlies successful performance in these tasks? The answers tend to revolve around notions such as *pre-emptive self-criticism* (Tetlock and Kim, 1987), *cognitive effort* (Kennedy, 1993), *thoroughness* (Brtek and Motowidlo, 2002), and *systematic, effortful processing* (de Langhe et al., 2011). Notions such as *imagination*, *insight*, and *originality*, on the other hand, tend to be missing. In this literature, little attention has been devoted to the effects of accountability on *creative problem solving*, or the ability to find an original and useful solution to a problem (Stein, 1953; Runco and Jaeger, 2012). Among the studies investigating process and outcome accountability, for example, only the one by Häusser et al. (2017) runs counter to the norm by investigating idea generation in a product development task. Given the prevalence of accountability in the modern workplace and the importance of creative cognition as a source of competitive advantage, this is an important omission.

Drawing on more than two decades of investigating creativity in organizations, Amabile (1998) suggests that although most managers recognize its value, "creativity gets killed much more often than it gets supported," and that this can happen when "new ideas are met not with open minds but with time-consuming layers of evaluation—or even with harsh criticism." Accordingly, she cautions against a "a culture of evaluation" that "leads people to focus on the external rewards and punishments associated with their output." These observations resonate with those of the pioneering computer scientist Alan Kay quoted in Chapter 3, and a similar line of thinking can also be detected behind the development of the brainstorming technique, as evidenced by the emphasis on a lack of critical judgment in the initial stages of the process (Taylor et al., 1958). These observations all highlight the potential of accountability to have a detrimental effect on creativity. Accordingly, a research direction concerned with finding ways of incorporating accountability that are least harmful (or even beneficial) to creativity would be of significant interest to a management system designer.

On a micro level, a common obstacle standing in the way of creative achievement is thinking about the problem in a wrong way due to an inappropriate *cognitive representation* of the problem space. Exemplars of such inappropriate cognitive representations include assuming a constraint that is not actually present and having too narrow a conception of the possible uses of some object or resource. As a result, initial problem-solving attempts are unproductive and often lead to a period of *impasse* during which the problem solver is stuck and does not know how to move forward. Sometimes, however, this impasse is punctured by a sudden and unanticipated grasping of the solution—the "Aha!" experience. Such an *insight sequence* appears to characterize a number of significant inventions and discoveries in business, science, and art. A classic illustration involves the Greek polymath Archimedes discovering a solution to a problem posed by the king while taking a bath and subsequently running naked through the streets of Syracuse shouting "Eureka!" To investigate insight in laboratory conditions, psychologists have collected and constructed a variety of toy problems that tend to be challenging because they invoke a misleading cognitive representation or that otherwise invite

incorrect solution attempts (Chu and MacGregor, 2011). When individuals do solve these problems, they often report having done so with a sudden insight. Accordingly, the label *insight problems* is typically used to collectively describe these problems. Ideally, a systematic study of experimental variables that affect the solution rate or the average time to solution for insight problems can yield useful clues as to what approaches could be used to stimulate useful and original solutions to the much more complex and messy problems faced by modern organizations.

So how can fruitful insights be facilitated? The standard solution in the face of a difficult problem is to think harder. Translated into the language of modern psychology, thinking harder typically involves an increase in *cognitive control* which enables us to tune out distractions and fully devote attentional resources to the current goal (Cohen, 2017). But is thinking harder always the right approach? In their review, Amer et al. (2016) draw on a methodologically diverse body of research to argue that cognitive control can actually be a "double-edged sword." That is, in some tasks and contexts—particularly those relating to creativity and insight—reduced cognitive control sometimes appears to result in better performance. A study of neurological patients provides an interesting example (Reverberi et al., 2005). In this study, individuals with focal damage to prefrontal cortex (part of the brain associated with cognitive control) were considerably more likely than matched healthy individuals to solve a difficult version of a matchstick arithmetic problem—of the kind that has often been used as a model in research on insight. Altogether, the research reviewed by Amer et al. (2016) suggests that increased cognitive control can occasionally be detrimental in tasks calling for insight and creativity.

In the accountabilty literature, cognitive control and the closely related concept of Type 2 processing are more strongly associated with process accountability than with outcome accountability, although—as noted before—the evidence for this claim is not as strong as one would hope. Nevertheless, the existing evidence raises the possibility that outcome accountability would prove more appropriate than process accountability in the context of insight problems. Such a prediction is in line with theoretical accounts of insight problem solving that emphasize the role of implicit, Type 1 processing in restructuring and insight (Ball et al., 2015). It is also in line with the suggestion that "the key to creativity is giving people autonomy concerning the means—that is, concerning the process—but not necessarily the ends" (Amabile, 1998). Finally, highly relevant evidence was reported by Eskenazi (2015) who reported a negative effect of process accountability on performance on remote associates problems—another problem type often used in research on insight. His studies, however, contrasted the process accountability condition with a no accountability condition, leaving open the possibility that the observed results were due to a more general accountability effect.

In Chapter 3, we study the effectiveness of process accountability and outcome accountability in facilitating performance on the two kinds of insight problems mentioned above—matchstick arithmetic problems and remote associates problems—while also aiming to obtain mediational evidence using self-report measures of Type 2 thinking and anxiety. Overall, by extending the study of process and outcome accountability to the domain of insight problems, the research reported in this chapter aims to provide a management system designer with a more complete picture of the relative desirability of these two forms of accountability.

1.6 Rationality and accountability

Research summarized in this chapter suggests that accountability, if implemented well, can serve a management system designer as an antidote to some of the irrational tendencies of the human mind. This seems like excellent news since the literature in psychology and behavioral economics volunteers plenty of evidence of human irrationality (Ariely, 2008). People's judgment and choice appear to rely on simple heuristics and systematically deviate from normative standards, such as rules of logic, probability, and rational choice (Kahneman, 2011). For example, unchecked intuitions can lead to errors on otherwise rather simple arithmetic tasks (Frederick, 2005) and probabilistic judgments can be insensitive to prior probabilities whilst being influenced by transparently irrelevant information (Tversky and Kahneman, 1974). These and many other cognitive biases appear to be widespread and often difficult to combat. For example, in their review of the literature on the effect of financial incentives, Camerer and Hogarth (1999) noted that "*no* replicated study has made rationality violations disappear purely by raising incentives." The suggestion that "debiasing people against errors in thinking could be among psychology's most enduring legacies to the promotion of human welfare" therefore does not come as a surprise (Lilienfeld et al., 2009). The many documented shortcomings of the human mind in its typical mode of operation seem to provide a compelling justification for accountability and other forms of influence.[7]

But another perspective on rationality emerged in recent decades from the research program on *ecological rationality* and *fast-and-frugal heuristics* (Gigerenzer and Todd, 1999; Todd and Gigerenzer, 2007; Todd and Brighton, 2016; Artinger et al., 2015). Researchers working in this area are skeptical of applying normative standards designed for "small worlds" of perfect knowledge and tractability to the uncertain and computationally intractable "large worlds" individuals typically operate in. Accordingly, they are also wary of relying on the observed violations of these normative standards to make claims such as the one that concludes the preceding paragraph. A key question for these researchers is if these violations are actually materially costly. For example, Arkes et al. (2016) conducted a systematic literature search in an attempt to uncover demonstrations of costly violations of coherence-based rationality norms (specifically, *transitivity*, *procedural invariance*, *descriptive invariance*, and *independence of irrelevant alternatives*) and concluded that "there is little empirical evidence that violations of coherence norms are costly, or if they are, that they survive arbitrage and learning."

As the name suggests, the research program on ecological rationality is built around an alternative notion of rationality that combines an *instrumental* and an *ecological* component. The basic question is, How can individuals[8] best achieve their goals (the instrumental component) given the external environment in which they operate (the ecological component)? Accordingly, the research strategy consists of two major tasks. The first task is identifying strategies that could serve specific goals and are at the same time psychologically realistic (exploiting evolved

[7] In addition, accountability can also be motivated from the classic economic perspective. That is, an accountable individual could be conceived as a rational agent whose interests are not sufficiently aligned with the interests of the principal. In this setting, accountability serves as one of the mechanisms that can facilitate the desired alignment by increasing monitoring and imposing novel potential consequences (Pitesa and Thau, 2013).

[8] I use "individuals" here because this is the focus of this dissertation. The scope of the ecological rationality program is broader as it includes groups, companies, cats, etc.

capacities and taking into account basic cognitive limitations). The second task consists of a systematic study of the conditions that determine how well these strategies perform in the relevant environment (Todd and Brighton, 2016).

A major discovery of this research program has been the existence of *less-is-more effects* (Gigerenzer and Brighton, 2009). Simple decision strategies that rely on less information and use less computation can in fact perform surprisingly well in comparison with strategies that use more of both. The *recognition heuristic* is the simplest example of such a *fast-and-frugal heuristic* that has been extensively studied (Goldstein and Gigerenzer, 2002). If an individual is asked to predict which of two cities—Rotterdam or Ljubljana—is larger and answers "Rotterdam" because it is the only one that she recognizes, she is relying on the recognition heuristic. This strategy is information-frugal because it uses only a single cue (recognition). In spite of this, its predictions can be highly accurate under the right conditions. The result of its *ecological rationality* when the correlation between recognition and criterion is positive and substantial is that less knowledgeable individuals can be more accurate than more knowledgeable ones on comparison tasks such as the city-comparison task illustrated above. In one demonstration, Londoners were more likely than New Yorkers to correctly guess which of two American cities—San Diego or Milwaukee—has the larger population (see Gigerenzer and Goldstein, 2011).

Other examples of fast-and-frugal heuristics that have proven surprisingly effective include the *Take The Best* heuristic (Gigerenzer and Goldstein, 1996), the *fluency* heuristic (Schooler and Hertwig, 2005), equal-weighting strategies such as the *1/N heuristic* (DeMiguel et al., 2009) and *tallying* (Dawes, 1979), the *hiatus heuristic* for classifying customers as active or inactive (Wübben and Wangenheim, 2008), *fast-and-frugal decision trees* (Woike et al., 2017), and Δ-inference (Luan et al., 2019). The Take The Best heuristic especially has been at the center of much empirical and theoretical work (e.g. Gigerenzer and Goldstein, 1996; Bröder, 2003; Rieskamp and Otto, 2006; Dieckmann and Rieskamp, 2007; Hogarth and Karelaia, 2007; Walsh and Gluck, 2016). It features a noncompensatory approach to multiple-cue decision making: the single highest-validity cue that discriminates between the two alternatives is used to make a decision. Despite its simplicity, simulations based on the data from many real-world decision environments show that Take The Best is often just as accurate or even more accurate than much more informationally greedy and computationally intensive strategies such as the *Weighted Additive* strategy (described below) in genuine prediction on new data (Czerlinski et al., 1999; Brighton, 2006; Gigerenzer and Brighton, 2009; Şimşek, 2013; Lee et al., 2017). A key takeaway from these and related findings is that the rationale for the use of heuristic strategies stretches beyond the idea of an *accuracy/effort trade-off* (acccording to which the use of heuristics can be justified by their decreased use of resources such as time and cognitive effort). As Artinger et al. (2015) put it when discussing the relevance of these results to management, "heuristics allow for adaptive responses to the characteristics of an uncertain managerial environment."

The research program on ecological rationality and fast-and-frugal heuristics provides a valuable theoretical perspective. It sensitizes the management system designer to the possibility that seemingly irrational or otherwise undesirable patterns of thought and behavior could in fact be free of meaningful cost (economic or otherwise) and therefore do not call for an elaborate accountability intervention. Furthermore, it demonstrates that simple heuristic strategies

that ignore some of the information can be not only efficient but actually more effective than the more complicated and seemingly more justifiable decision strategies. Accordingly, Artinger et al. (2015) stress "the importance of management practitioners being aware of the different basic decision strategies that have been identified as successful tools in an uncertain environment." This knowledge is highly pertinent to the task of the management system designer, not least because many conditions of high accountability could plausibly select against these kind of simple and effective strategies. For example, Hogarth (2012) suggests that "in the litigious environment of medicine, simple decision rules that ignore information may be resisted." These observations increase in importance when one considers the costs that accountability systems neccessarily impose. Consider, for example, a choice concerning how much should the evaluator monitor the performance of the accountable individual. An obvious cost of increasing the degree of scrutiny is the time cost of the evaluator. But if the system involves a competitive component (for example, it results in one employee receiving a performance bonus), Kareev and Avrahami (2007) demonstrate a less obvious cost. Although a high degree of scrutiny yields more accurate assessment, it can also result in lower *overall performance* than a lower degree of scrutiny. The basic insight is that a *lower* scrutiny increases sampling error and therefore increases the chance that a lower-performing indvidual will be rewarded which in turn increases the motivation of both lower-performing and higher-performing individuals. In accountability, too, less can be more.

The ideas and methods of the research program on ecological rationality and fast-and-frugal heuristics inspired two of the chapters in this dissertation. The first of these is Chapter 4. Building on the idea that effective decision making consists of the ability to select a decision strategy that exploits the structure of the environment, we examine how process and outcome accountability serve this goal in two kinds of decision environments. One of these environments resembles those used in several previous studies on process and outcome accountability. In this environment, which we call the *compensatory environment*, the use of the information-frugal Take The Best heuristic leads to relatively low decision accuracy. In contrast, using the information-intensive Weighted Additive strategy that integrates all available information leads to high accuracy. The Weighted Additive strategy requires that all decision cues be inspected and integrated in a weigh-and-add manner typical of classical normative theories of choice. In addition to the compensatory environment, we investigate a *noncompensatory environment* which is its mirror image in this crucial aspect. Meaning, the use of Take The Best heuristic leads to high decision accuracy whereas the use of the Weighted Additive strategy leads to low decision accuracy. Importantly, this crucial difference between the two environments is not immediately obvious but must be inferred on the basis of accuracy feedback. Accordingly, we use the term *adaptive decision making* to describe the subjects' task. While the primary objective of this research is to investigate how well process- and outcome-accountable individuals adapt to the two kinds of decision environments, the secondary objective is to investigate the underlying cognitive processes through the use of self-report and process-tracing techniques.

The research program on fast-and-frugal heuristics also inspired Chapter 5, which makes a more direct contribution to this literature. Building on the observation that fast-and-frugal heuristics can be ecologically rational, a few previous studies investigated their use among experienced decision makers. Contradicting the assumption that expertise involves a relatively

complete and compensatory use of information, the results of these studies suggest that the majority of experienced decision makers rely on fast-and-frugal heuristics (Garcia-Retamero and Dhami, 2009; Pachur and Marinello, 2013). These studies have, however, considered decision environments characterized by a low degree of accountability. For example, Pachur and Marinello (2013) investigated airport patrol officers who make hundreds of quick decisions each day concerning which passengers to stop and search. To complement these preliminary findings, we investigate a very different decision environment, that of professional loan officers. Commercial lending presents an interesting case, as the decision making is relatively slow, the stakes are high, and the need to justify one's decisions looms large. These conditions may facilitate the kind of information-intensive decision making that is modeled by the Weighted Additive decision strategy (Lee et al., 1999). On the other hand, the business environment is highly uncertain and changing, and fast-and-frugal heuristics are thought to be particularly effective in these kinds of environments (Artinger et al., 2015). Accordingly, we might expect loan officers to adapt to these conditions by relying on heuristics. To investigate which of these perspectives predicts the data better, we conducted a decision-making experiment featuring professional loan officers and used Bayesian cognitive modeling to individually classify each loan officer to one of three common decision strategies. In addition to reporting these results, the chapter makes two other contributions. First, a theoretical contribution is the development of a new variant of the information-intensive Weighted Additive strategy that incorporates a more psychologically realistic theory of error, based on the idea that errors (deviations from what the strategy predicts) should become less likely as the overall value difference between the alternatives increases. Second, an empirical contribution is the use of eye-tracking technology to trace cognitive processes preceding choice and thus provide further insight into loan officers' decision making.

1.7 Accountability in the lab and beyond

This dissertation and much of the experimental research on accountability is fundamentally motivated by the goal of helping management system designers in incentivizing thinking and behavior that is likely to lead to desirable outcomes. That is, the ultimate goal is to generate knowledge that provides useful input to the designer's decision making on matters of accountability—especially knowledge that would otherwise be hard to come by. In the service of this goal, laboratory researchers can leverage their freedom to set up (un)usual accountability arrangements and carefully trace their effects. But this freedom comes at a price. Even if researchers motivate some design choices by referring to features of organizational life, many other aspects remain highly artificial by neccessity or convenience. In addition, the subjects in these studies tend to be students rather than experienced professionals. In the case of experimental accountability research, Aleksovska et al. (2019) estimate that 73% of studies published in the past half-century relied on student samples. Accordingly, there is typically a considerable "distance" between the laboratory setting in which a theory has been tested and the specific real-world setting of the management system designer. The latter may justifiably wonder how much faith to place in laboratory results in light of this fact.

The basic problem is that an experiment never tests the theory alone. What gets tested is

the union of theory and auxiliary assumptions. That is, a large and mostly implicit set of assumptions is required to lower the theoretical concepts down the ladder of abstraction into, eventually, the words printed on the paper, displayed on the screen, or spoken by the researcher; other symbols shown to the subject; other behaviors of the researcher; the laboratory coat (not) worn by the researcher; the audio recorder (not) in full view of the subject—and everything else that forms the total stimulus package encountered by the subject, including the color of laboratory walls. Accordingly, a question arises: How much of its apparent predictive success is actually due to the theory and how much due to a favorable constellation of auxiliary assumptions (arising from e.g. luck or the researcher's intuition)? In other words, how robust is the predictive power of the theory to variation in those aspects not explicitly included in the theory?

An elegant solution to this problem is to formulate theories that make precise and (outside of the theory) unlikely predictions (Popper, 2005). For even a single laboratory corroboration of such predictions can be highly informative. By way of analogy, imagine that a stranger tries to sell you a machine that—he claims—has some predictive power when it comes to lotteries and he agrees to lend it to you for a test-drive with a local lottery that is scheduled soon. The winning sequence in this lottery is composed of integers 0—9. On the day of the lottery, you adjust the knobs on the machine to input the information it needs and the machine dutifully outputs its prediction, which is then confirmed in the evening by the lottery draw. Now, if the entire prediction was that the first digit of the winning sequence would be 4, you would be mildly impressed; if the machine was junk, you would expect its chance of such a success to be only 10%. If, instead, the machine correctly predicted the first two digits—4, 2—your enthusiasm for the machine would be considerably greater. And so on: the more digits the machine predicts, the more convinced you are that the predictive success really has to do with the theory of lottery built into the machine. Analogously, the more precise and otherwise unexpected the theory's predictions that get corroborated by the laboratory experiment, the more justified is the management system designer in placing faith in the theory.

Another solution involves the simple act of repetition. To return to the lottery analogy, if the machine only predicts the first digit but the stranger agrees to let you use the machine for the next two, three, or more iterations of the local lottery, you are in a better position to assess its value than if you only get the one test. Similarly, in laboratory research investigators can repeatedly test a theory to see how robust it is. If a theory (be it an elaborate one or a mini-theory in the form of an "effect") proves a reliable guide even when tested by independent investigators, the management system designer may wish to take notice. But if a theory is not able to demonstrate reliable predictive success even in laboratory settings, what is its use in real-world settings? *Replication studies* are an excellent tool to reveal gaps in current theories and methods. These gaps may be especially revealing and hard to deny when they result from replication studies that aim to mimic the original study as closely as possible. Conducting replication studies therefore plays an important role in driving the cumulative progress of a field of research (Zwaan et al., 2018).

Importantly, the benefit of repeated testing is undermined if only a biased subset of all results emerges in the published record. Returning to the lottery analogy, if the stranger agrees to report to you several tests for his machine but does not commit to reporting the outcomes of *all* tests, you'd be justifiably skeptical of the reported track record. Similarly, the existence of

the *publication bias* in the academic literature makes it more difficult to assess the empirical status of a theoretical claim (Greenwald, 1975; Smaldino and McElreath, 2016). When the incentives for reporting a certain kind of outcome (for example, a statistically significant result that is consistent with previous published research) are paired with substantial freedom in conducting, analyzing, and reporting of research, the published record becomes even less credible (Simmons et al., 2011; John et al., 2012). Analogously, if the stranger not only decided whether to report a given test result or not but could additionally pick, after the fact, one of several predictions the machine had made, the reported successes would be even less impressive. These considerations make it all the more important to conduct and publish replication studies (Zwaan et al., 2018).

What does this mixture of observation and conjecture imply for experimental laboratory research on accountability? In the following I formulate five recommendations for future research and, where applicable, highlight the associated contributions of this dissertation.

First, we should try to formulate theories that generate multiple, precise, and—in the absence of the theory—unexpected predictions. Second, we ought to test as many of these predictions as feasible. For example, Dalla Via et al. (2018) predicted that the positive effect of process (versus outcome) accountability on decision quality would be mediated by *search effort* and collected eye-tracking data to conduct a fuller test of the theory. In contrast, Siegel-Jacobs and Yates (1996) argued that outcome (versus process) accountability results in greater *stress* but neither they, nor any other paper that referred to this possibility, tested the hypothesis—until, twenty years later, Häusser et al. (2017) finally did (and found no evidence for it). In the studies reported in this dissertation, the tests of predictions concerning performance/accuracy are complemented with tests of the associated psychological processes, using data from self-report and process-tracing measures.

Third, more replication studies should be conducted and published since, especially in the literature on process and outcome accountability, independent replication studies are virtually nonexistent. Novel publishing options such as *registered reports* (providing in-principle acceptance of a paper before the data are collected; Chambers, 2013) can help ensure that well-done replications do not end up in a "file-drawer" (Rosenthal, 1979). In this dissertation, replication takes center stage in Chapter 2, where we discuss its importance and attempt to replicate a few key results from the literature on process and outcome acccountability. Another form of replication that should become more common is reuse of measures that have shown promise in explaining accountability effects. For example, the situation-specific scale of rational thinking style is validated (Novak and Hoffman, 2009), short, can be administered as part of the postexperimental questionnaire, relates to current theorizing on the effects of process accountability, and has been shown to mediate the performance difference between process and outcome accountability in one previous study (de Langhe et al., 2011). If, for example, Dalla Via et al. (2018) included the measure in their post-experimental questionnaire, they could have replicated the previous results and compared the explanatory power of this measure with their main process measure based on eye-tracking. In our research, the scale of Novak and Hoffman (2009) was included in four studies on process and outcome accountability; the results are presented separately in the associated chapters and jointly in Chapter 6.

Fourth, to accelerate the conduct of replication studies (among other benefits), sharing of

experimental materials (all of them, but especially those concerning the manipulations of accountability) should become the norm. Fifth, routine sharing of data should also become the norm, especially in the absence of *preregistration* (Nosek et al., 2019) or *multiverse analysis* (Steegen et al., 2016). Openly shared data allows other interested researchers to check if the reported statistical results are robust to alterrnative—but still reasonable—analytical decisions. Accordingly, the experimental materials and data of the studies reported in this dissertation are publicly archived on the Open Science Framework website (the URLs are provided at the end of each Methods section).

I submit that following these recommendations would enable faster cumulative progress of accountability research and, over the long run, increase its usefulness to management system designers.

1.8 Declaration of contribution

The following text declares my contribution to each of the chapters in this dissertation and acknowledges the main contributions of others. Additional acknowledgments are provided on the first page of each chapter and in the *Acknowledgments*.

Chapter 1: I wrote the chapter. Frank Hartmann and Maarten Boksem provided helpful suggestions following the initial outline as well as feedback on the completed text.

Chapter 2: I developed the research questions, reviewed the literature, designed the study and prepared the materials, collected and analyzed the data, interpreted the results, and wrote the chapter. Frank and Maarten provided helpful suggestions and critical feedback every step of the way.

Chapter 3: I developed the research questions, reviewed the literature, designed the studies and prepared the materials, collected and analyzed the data, interpreted the results, and wrote the chapter. Frank and Maarten provided helpful suggestions and critical feedback every step of the way. The first study was designed together with Franziska Grotheer who also collected the majority of the data and wrote her Master thesis on the basis of the study.

Chapter 4: I developed the research questions, reviewed the literature, designed the studies and prepared the materials, collected and analyzed the data, interpreted the results, and wrote the chapter. Frank and Maarten provided helpful suggestions and critical feedback every step of the way.

Chapter 5: I developed the research questions, reviewed the literature, designed the studies and prepared the materials, collected and analyzed the data, interpreted the results, and wrote the chapter. Frank and Maarten provided helpful suggestions and critical feedback every step of the way.

Chapter 6: I wrote the chapter. Frank Hartmann and Maarten Boksem provided helpful suggestions following the initial outline as well as feedback on the completed text.

Chapter 2

On multiple-cue learning under process and outcome accountability

Steer your heart past the Truth you believed in yesterday
Such as Fundamental Goodness and the Wisdom of the Way
— Leonard Cohen takes replication studies very seriously

Summary:

The existing literature on process and outcome accountability is characterized by a lack of independent replication studies. In this chapter we highlight the importance of replication and contribute to this important scientific goal by replicating and extending two key findings in the accountability literature, originally reported by de Langhe et al. (2011). In particular, we compare objective performance and self-reported thinking style of process- and outcome-accountable subjects in a multiple cue learning task. Besides replicating the original version of the task that revealed superior performance under process accountability, we investigate a version of the task that includes uninformatives cues as well as novel measures of task knowledge, decision time, and (over)confidence. In line with the original research, we obtain some indication that performance and rational thinking style increase under process accountability relative to outcome accountability. At the same time, the statistical evidence in the current study is much more equivocal and the magnitude of these effects is about half of those in the original study. Additional analyses provide some evidence that process accountability increases explicit task knowledge and possibly decision time compared to outcome accountability, whereas the degree of calibration of confidence judgments is similar in the two conditions. We conclude by providing recommendations for future original and replication research on process and outcome accountability.

Acknowledgments:

We thank Bart de Langhe for providing us with information on the experiments reported in de Langhe et al. (2011) and for reviewing our experimental materials.

2.1 Introduction

In psychology and management, recent years have witnessed a growing interest in the effects of process and outcome accountability on human judgment and decision making (Patil et al., 2014). Whereas process-accountable individuals are asked to justify their decision-making approach and held accountable for its quality, outcome-accountable individuals are held accountable for the eventual outcomes of their decisions (Siegel-Jacobs and Yates, 1996). The distinction has proven useful in explaining some of the variation in the effects of accountability on performance and well-being (Hall et al., 2017) and the relative merits of each accountability type are often debated in fields ranging from education to intelligence analysis (Tetlock et al., 2013). In laboratory research, the focus has been primarily on contrasting the two types of accountability in a variety of domains (Patil et al., 2014). Some research has additionally begun to elucidate the cognitive and affective processes that underlie the observed differences (e.g. de Langhe et al., 2011; Eskenazi, 2015; Häusser et al., 2017; Dalla Via et al., 2018). What has been notably missing, however, are independent replication studies.

Replication studies are a cornerstone of the scientific method. They separate robust phenomena that are useful targets for further study and theorizing from "occult effects" (Popper, 2005) that are not. Although traditionally neglected and difficult to publish in psychology and many other social sciences, replication studies have gained in prominence in recent years (Zwaan et al., 2018). For example, several recent large-scale projects had replication as their central aim (e.g. Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Klein et al., 2018). In psychology, the most influential of these has been the effort to replicate a set of 100 findings published in the major journals of the field (Open Science Collaboration, 2015). The poor replication rates reported by these projects drive home the point that efforts to discover new phenomena need to be accompanied by efforts to replicate published claims. As Gigerenzer (2018) recently noted, "science is a cumulative endeavor, not a yes/no decision based on a single empirical study."

Replication studies can be conducted by researchers who originally discovered the finding of interest or by independent investigators. Replication studies conducted by the latter are called *independent replications* and are typically more valuable than replications conducted by the original researchers, for two reasons. First, each finding can be expressed as a function of factors (relating to instructions, manipulations, tasks, measures, subjects, etc.) that are theoretically important and those that are not (without this distinction, experimental research becomes a purely historical discipline and loses its *raison d'être*). But any given finding may be crucially dependent on factors that are at first thought to be theoretically unimportant. Independent investigators are less likely to replicate these factors in their attempts and their results are therefore more likely to reveal the gap between theory and truth. The second reason to privilege independent replications is psychological; independent investigators are typically less invested in a given finding and thus more likely to be unbiased in their conduct and interpretation.

In the literature on process and outcome accountability, replications are relatively rare. And if we are talking about *independent* replications, they seem to be almost nonexistent[1]. This is

[1] But see Schulz-Hardt et al. (2020); we discuss this work in Chapter 6.

problematic because there is arguably a greater need for replication in accountability research than in many other areas of experimental psychology. This claim is based on two observations. First, the distinction between process and outcome accountability is conceptually broad, providing only light guidance and constraint for measurement and operationalization. Researchers are thus guided by their own idiosyncratic judgment when operationalizing process and outcome accountability. As a result, we observe considerable variation in experimental materials used to manipulate these two types of accountability, and it is unclear to what extent these variations tap into a common core. This issue is amplified if, as Patil et al. (2014) argue, relational micro-signals—possibly substantiated merely in "a shift in one or two words"—can have "big effects on the cognitive, emotional, or behavioral responses of agents to accountability guidance." Replication helps reveal the (lack of) robustness of reported findings and facilitates the discovery of factors that drive the observed phenomena. This facilitates theory development and makes it easier to formulate useful managerial recommendations.

The second (and closely related) observation is that manipulations of accountability are inherently social. As a result, even when the same experimental materials are used, there may be meaningful variation across studies in how subjects respond to these materials. In accountability studies, the behavior of the experimenter and the broader context (for example, the institution at which the research is taking place) could plausibly influence the behavior of subjects. For example, earlier accountability research shows that when the views of the audience are known, there is a tendency to conform to these views (e.g. Tetlock et al., 1989). A process of this kind may operate even when the views of the audience are not explicitly manipulated and recorded as a part of the stimulus package. As a result, certain findings may be highly context-dependent.

The concerns grounded in these observations map onto two kinds of replication studies. First, the concern that variation in experimental materials may be consequential is best addressed by so-called *conceptual* replications (Hüffmeier et al., 2016). In a conceptual replication, one or more aspects of the original study are changed. For example, a different experimental manipulation may be used as the independent variable or a different measure may be used as the dependent variable. If the finding persists across this variation, our confidence in its generalizability and importance is increased. Second, the concern that context effects (that are not captured by recorded experimental materials) may be consequential is best addressed by so-called *close* replications (Hüffmeier et al., 2016). In a close replication, an attempt is made to match the characteristics of the original study as closely as possible. Successful close replications assure us that a given finding is not an "occult effect" (Popper, 2005) but that it can instead be reliably instantiated by following a particular experimental "recipe." Although the distinction between conceptual and close replications can be challenged on the grounds that the distinction is not sharp (Gelman, 2018), we believe it can serve a useful role in communication. The work reported in this chapter features elements of both close and conceptual replication. In the framework of Hüffmeier et al. (2016), it would be characterized as a *constructive replication* because it features a condition that closely matches the original study as well as a condition that "adds new elements" (the constructive condition) and can therefore "extend or refine" the basic finding.

In the current research, we primarily aimed to replicate and extend two key findings in the accountability literature. These findings were originally reported by de Langhe et al. (2011),

who investigated the relative effects of process and outcome accountability using a *multiple-cue learning task*. Subjects were asked to predict a criterion on the basis of several cues and received outcome feedback that could be used to improve performance. In particular, their task was to predict the popularity of a specific category of consumer products (*EasyPhones*) on the basis of a few visual cues (such as color and shape). Tasks of these sorts are used to simulate experiential learning and Matton et al. (2013) provide some evidence of their external validity in predicting performance among pilot trainees. In their studies, de Langhe et al. (2011) implemented two different versions of their task. In the *elemental* task the cues were linearly and additively related to the criterion whereas in the *configural* task the relationship between cues and criterion was nonlinear. Whereas *cue abstraction* can be used to successfully complete the former, the latter is more difficult and calls for *exemplar processing*. The principal finding of de Langhe et al. (2011) was that accuracy in the configural task was very similar in the two accountability conditions whereas in the elemental task process-accountable subjects outperformed outcome-accountable ones. Moreover, de Langhe et al. (2011) used the situation-specific rational thinking style questionnaire (Novak and Hoffman, 2009) to provide evidence concerning the underlying mechanism. Relative to outcome accountability, subjects in the process accountability condition scored higher on this questionnaire. In the language of dual-process theories, this suggests that process accountability increases reliance on Type 2 processes, i.e. controlled processes mediated by working memory (Kahneman, 2011).

Our first aim in this research was to conduct a close replication of the finding of superior performance of process-accountable subjects in an elemental task. In the current study, we refer to this version of the task as the *standard environment*. Second, we considered a variation of the task which we refer to as the *mixed environment*. Whereas all cues in the standard environment are predictive of the criterion, some of the cues in the mixed environment are completely uninformative. This version of the task can be easily motivated with reference to the real world, which is under no obligation to only serve us useful information. Information that is not predictive can nevertheless influence judgments. One demonstration of this is found in the research on the *dilution effect* which shows how adding nondiagnostic information to diagnostic information can decrease judgment quality (Nisbett et al., 1981). Interestingly, whereas process accountability typically has positive effects on judgment quality, there is some evidence that it may exacerbate this tendency (Tetlock and Boettger, 1989; but see Tetlock et al., 1996; Hoffman and Patton, 1997). The inclusion of the mixed environment was additionally motivated by some preliminary evidence we obtained that outcome-accountable subjects may adapt more quickly than process-accountable ones to decision-making environments that facilitate the use of information-frugal strategies (which ignore some of the available information; see Chapter 4).

Our second aim was to replicate the finding that process accountability (relative to outcome accountability) increases situation-specific *rational thinking style* as captured by the self-report instrument of Novak and Hoffman (2009). This finding is important because it provides evidence regarding the underlying mechanism. Our interest in replicating this finding was further bolstered by a failure to replicate it in a different study (see Chapter 4[2]). In that study, the mean level of rational thinking as measured by the same instrument was essentially the same under process and outcome accountability. But because thinking style is partly a function

[2] This comment refers to the preliminary study summarized in the section *Introduction*.

of the task (Novak and Hoffman, 2009; Söllner et al., 2013), the different task we employed serves as one plausible candidate explanation of our failure to replicate. In the current study, we aimed to recreate the conditions that produced the original result as closely as possible.

The final aim of the current research was to extend the work of de Langhe et al. (2011) by including the measures of explicit task knowledge, decision time, and subjective confidence, therefore providing further insight into the cognitive processes underlying the differences between process and outcome accountability. The first two of these measures can provide converging evidence that the superior accuracy of process accountability is mediated by Type 2 processes, whereas subjective confidence can be used to investigate metacognitive processes and the degree of *calibration*. In one of the earlier studies on process and outcome accountability, Siegel-Jacobs and Yates (1996) found evidence for better calibrated probability judgments among process-accountable subjects versus outcome-accountable subjects.

LeBel et al. (2019) note that while the importance of replication studies is increasingly recognized in psychology, there is less consensus on how these studies ought to be evaluated. Addressing this gap, the authors propose a two-stage approach. One of these stages concerns the way the results are summarized and is therefore taken up in the section *Discussion*. The other stage involves six crucial study characteristics that should be used to describe and evaluate a replication attempt. The first two of these concern the similarities and differences between the original study and the replication. In the current study, the comparison of process and outcome accountability in the *standard* environment undoubtedly constitutes a close replication. Care was taken to implement the condition in the same way de Langhe et al. (2011) did, and the subjects were sampled from the same subject pool[3]. For the *mixed* environment condition, the label of a "constructive replication condition" proposed by Hüffmeier et al. (2016) seems best. As noted above, our inclusion of fully uninformative cues in this condition could turn out irrelevant when it comes to the main findings of interest. On the other hand, it could have a moderating role, similar to how the switch from an elemental to a configural task turned out to be meaningful in the original research (de Langhe et al., 2011). In both the standard and the mixed environment, we collected some additional measures that were not collected in the original research. However, these differences were either undetectable from the subject's perspective (decision time) or occured—unannounced—*after* the main part of the task (e.g. the confidence probability question in Part 2 of the task or the explicit knowledge test). Apart from this, there are no meaningful differences between the current study and the original study. To ensure that this is the case, one of the authors of the original study was first consulted for details and later also reviewed the materials. Next, LeBel et al. (2019) highlight the issue of *investigator independence*, noting that original authors and those closely affiliated with them might be more susceptible to bias in conducting and evaluating the study. In the current study, this is not an issue. The next two characteristics concern the transparency of the study and the reproducibility of results. LeBel et al. (2019) emphasize that studies cannot be properly scrutinized if they are not reported with sufficient transparency. To address this issue, we openly share the materials, data, and analysis code and use freely available, open-source software to both conduct the experiment and analyze the results. Finally, LeBel et al. (2019) urge researchers to consider the relevant *auxiliary hypotheses*. For example, a study might

[3]Of course, this does not license the claim that the two samples belong to the same population; the characteristics of students taking part in the subject pool may have changed over time, for example.

fail to replicate an earlier finding because of a faulty measuring instrument. In the current study, we have added a manipulation check for accountability type that was not included in the original research.

2.2 Methods

Subjects

We recruited 192 subjects (mostly business students) from a paid subject pool. The average age was 21.7 years (SD = 2.4), 126 subjects were female, 67 male. In addition to the baseline payment of 10€, some subjects received a performance-contingent bonus (see *Experimental manipulations*). Data collection took place between March and June 2019. We initially planned to collect data from 200 subjects. But as subjects stopped signing up for the sessions and because the actual sample size was very close to the target, we stopped data collection. This decision was made before the data were analyzed.

Experimental design

Subjects were randomly assigned to one of the four experimental conditions obtained by crossing the type of accountability (process, outcome) and the type of environment (standard, mixed). There were 45 subjects in the condition process accountability × standard environment, 54 subjects in the condition outcome accountability × standard environment, 43 subjects in the condition process accountability × mixed environment, and 51 subjects in the condition outcome accountability × mixed environment.

Procedure and materials

Subjects (between one and eight per session) were jointly welcomed and escorted to individual cubicles[4]. After reading the informed consent statement and agreeing to take part in the study, subjects read the instructions (which included the accountability manipulation) and completed the main task. Afterwards, subjects completed the postexperimental questionnaire, which included manipulation check items, the rational (10 items) and experiential (10 items) subscales of the situation-specific thinking style questionnaire[5] (Novak and Hoffman, 2009), and a knowledge test. The manipulation check included two process accountability-related items ("I was informed that I will be asked to justify my prediction strategy when the prediction task is over" and "My final evaluation (and the chance to win up to 50 EURO if selected) is based only on how well I learned to justify my predictions") and two outcome accountability-related items ("I was informed that when the prediction task is over my overall accuracy will be computed by comparing my predictions with the actual popularity of EasyPhone models" and

[4] Subjects were asked to leave their phone outside before entering the cubicle. All but two subjects complied.
[5] We made slight adaptations to the wording of some of the items to make them suitable for our setting.

"My final evaluation (and the chance to win up to 50 EURO if selected) is based only on the accuracy of the predictions I made"). The knowledge test captured subjects' explicit knowledge of the relationship between cues and popularity. It consisted of two sets of questions. In the first set (*cue weights items*), subjects assessed the importance of each of the four cues (e.g. color). In the second set (*cue value items*), subjects were asked to directly compare the two possible values of each cue. For the *color* cue, for example, subjects indicated either that blue phones tended to be more popular, or that red phones tended to be more popular, or that there was no difference. Apart from this second set of items in the knowledge test, all questions in the postexperimental questionnaire were answered using a 5-point Likert scale. Subjects in the process accountability conditions were next asked to justify their approach. Finally, everyone entered their birthyear and gender. Apart from the informed consent statement (on paper), the procedure was fully computerized. The experiment was programmed in Python using the PsychoPy package (Peirce, 2007).

Main task

We used the phone popularity prediction task from de Langhe et al. (2011). Subjects assumed the role of a marketing manager tasked with predicting the popularity of EasyPhones, a type of phone designed specifically for the elderly. In each trial, they were presented with a picture of an EasyPhone (see Figure 2.1 for an example) and asked to predict its popularity on a scale from 0-10. The EasyPhones differed by *color* (*blue* or *red*), *size* (*large* or *small*), *shape* (*rectangular* or *curvy*), and *keyboard arrangement* (*plus-shaped* or *circle-shaped*). In the first 110 trials of the task, predictions were followed by immediate feedback about the actual popularity of the EasyPhone. In the final 32 trials, no such feedback was given. Instead, subjects were asked to indicate their subjective confidence that the prediction they provided was within ± 0.5 of the actual popularity. Formally, each trial was a judgment task in which subjects judged the *criterion* (popularity) of a single *object* (EasyPhone) described by four binary *cues* (color, size, shape, and keyboard arrangement). Therefore, there were in total 16 EasyPhones, defined by cue arrangements reproduced in Table 2.3 in section *Appendix 1*. Following de Langhe et al. (2011), we used a subset of 11 phones in the first part (first 110) trials and all 16 phones in the second part (last 32 trials). We randomized the order of the trials for each subject and each repetition of the subject, as well as the pairing of the observable EasyPhone characteristics (e.g. color) and the underlying cues (see *Experimental manipulations*).

Experimental manipulations

Four experimental conditions were created by crossing accountability type (process accountability, outcome accountability) and type of environment (standard, mixed environment). The manipulation of accountability type occurred right before the main task and was taken verbatim from Experiment 3 in de Langhe et al. (2011). Subjects were informed that their performance would be evaluated and that some (randomly selected) subjects would receive an additional payment (0€ - 50€) based on this evaluation. Subjects in the process accountability conditions were informed that one of the EasyPhones would be shown again, that they would

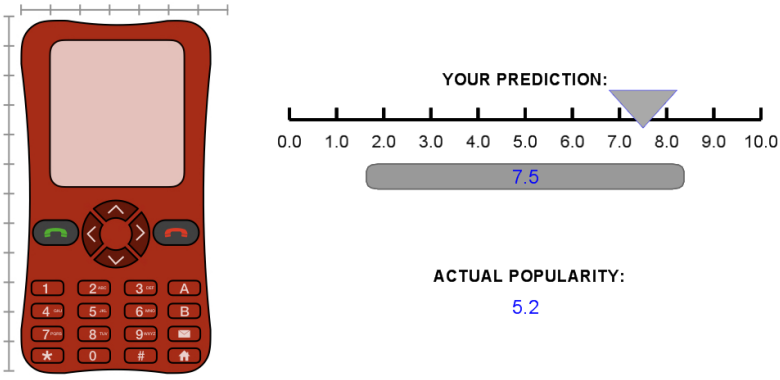


Figure 2.1: Screenshot of one trial in the main task. (Color version of this figure is available in the electronic version of this document.)

be asked to justify how they could determine its popularity, and that their evaluation would be based solely on the quality of their argument. In contrast, subjects in the outcome accountability conditions were informed that their accuracy would be computed and that they would be evaluated solely on the basis of accuracy. To manipulate the type of environment, we varied the weights of the four cues. The standard environment condition was a close replication of Experiment 3 in de Langhe et al. (2011), with the observed popularity of an EasyPhone i on trial j given by:

$$\text{Popularity}_{ij} = 4\text{Cue } 1_i + 3\text{Cue } 2_i + 2\text{Cue } 3_i + 1\text{Cue } 4_i + \varepsilon_{ij}. \tag{2.1}$$

In the mixed environment conditions, two of the cues were fully uninformative because their actual weight was zero. The popularity of the EasyPhones was thus determined by:

$$\text{Popularity}_{ij} = 2.4 + 4\text{Cue } 1_i + 2\text{Cue } 2_i + 0\text{Cue } 3_i + 0\text{Cue } 4_i + \varepsilon_{ij}. \tag{2.2}$$

The random error ε_{ij} was sampled from the same distribution in both environments:

$$\varepsilon_{ij} \sim \text{Uniform}(-0.5, 0.5). \tag{2.3}$$

Data analysis

We used Bayesian estimation to analyze the data. The models were fit using the R package `brms` (v2.4.0, Bürkner, 2017), which serves as an interface to the `Stan` language (Gelman et al., 2015). Our general strategy was to obtain a single value per subject for each measure by averaging over the relevant trials or items. We then regressed the resulting scores on accountability type,

environment type, their interaction, as well as the gender covariate. The latter was included because we observed gender differences for several measures of interest. We used sum-to-zero coding for accountability type (process accountability: +0.5, outcome accountability: -0.5), environment type (mixed environment: +0.5, standard environment: -0.5), and gender (male: +0.5, female -0.5). Depending on the nature of the dependent variable, we used different response distributions and link functions. If the latter are not explicitly identified for a given analysis, it means that the Gaussian distribution and an identity link function were used. To complete the models, we prepared informed priors using domain knowledge and previous results (de Langhe et al., 2011). It turned out, however, that for these data the default priors of the `brms` package result in essentially the same estimates as our custom priors, so we report the estimates based on default priors instead. When presenting the results in tabular form, we report posterior means and standard deviations. In addition, we use asterisks to highlight those parameters for which the posterior probability that the effect is in a particular direction (i.e. positive or negative) is greater than or equal to 0.95 (one asterisk), 0.975 (two asterisks) or 0.995 (three asterisks). Brackets next to the estimates in the main text provide 95% uncertainty intervals, i.e. [2.5th percentile, 97.5th percentile].

Readers uncomfortable with Bayesian statistics need not despair. For these models, the reported posterior means and standard deviations are almost exactly the same as classical linear regression estimates and standard errors. And the asterisks can be used to assess statistical significance. One asterisk translates to statistical significance at the 10% level (or 5% for a one-tailed test), two asterisks to statistical significance at the 5% level, and three asterisks to statistical significance at the 1% level.

Additional information

Additional information, including study materials and data, is available in the project's online repository (`https://osf.io/25egd/`).

2.3 Results

Preliminary analyses

Basic descriptive statistics are reported in Table 2.1 and Figure 2.2. The table reports, separately for each condition, the means and standard deviations of the measures used in the study, and the figure displays the correlations between the measures. The correlations make intuitive sense, as higher prediction accuracy correlates with spending more time on the predictions, engaging in a more rational thinking style, posessing more accurate explicit knowledge, and having greater confidence but lower overconfidence.

We formed a manipulation check for accountability type using the relevant items from the postexperimental questionnaire by subtracting the average of the two outcome accountability items from the average of the two process accountability items. As expected, the average score was positive in the process accountability conditions and negative in the outcome accountability

conditions. But we also found nine (4.6%) subjects with a manipulation check score that was directionally inconsistent with their accountability condition, and 29 (15.0%) subjects that had a neutral score of zero. We therefore conducted additional "compliant-only" analyses by excluding these subjects and rerunning all models. The results are reported in section *Appendix 3*.

Table 2.1: Condition means (standard deviations) for various measures, separately for each experimental condition (see section *Appendix 2* for definitions).

	Process Standard	Outcome Standard	Process Mixed	Outcome Mixed
Gender (% female)	66.7	59.3	69.8	66.7
Age (years)	21.9 (2.4)	21.7 (2.1)	21.6 (2.3)	21.8 (2.9)
Manipulation check	1.0 (1.1)	-1.9 (1.3)	1.4 (1.1)	-1.8 (1.3)
Decision time (seconds)	5.4 (2.1)	5.3 (3.9)	5.6 (2.4)	5.1 (1.7)
RMSE	2.0 (0.5)	2.2 (0.5)	1.6 (0.5)	1.7 (0.5)
Rational thinking	3.8 (0.7)	3.7 (0.6)	3.8 (0.6)	3.7 (0.7)
Experiential thinking	2.7 (0.7)	3.0 (0.8)	2.9 (0.8)	2.9 (0.9)
Knowledge	0.4 (0.6)	0.1 (0.8)	-0.1 (0.8)	-0.3 (0.9)
Confidence	63.0 (12.8)	61.9 (9.7)	65.3 (13.2)	65.1 (16.4)
Overconfidence	36.1 (15.2)	35.7 (12.3)	16.2 (25.3)	20.6 (20.9)
Brier score	0.3 (0.1)	0.3 (0.1)	0.3 (0.1)	0.2 (0.1)
Justification (word count)	69.8 (41.5)		65.2 (37.3)	
Justification (seconds)	216.5 (160.9)		207.7 (151.3)	

Note: Columns 2 - 4 correspond to the four experimental conditions defined by accountability type and environment type. For example, column 2 corresponds to the process accountability condition in the standard environment.

Predecisional behavior

We analyzed decision time, or the average number of seconds a subject took to enter their predictions. Having observed some large values that may reflect contaminant processes such as taking a break during the task, we computed the mean and standard deviation of each subject and excluded the 567 (2.1%) individual trials that fell three standard deviations above the mean. The remaining trials were averaged for each subject and modeled using the log-normal response distribution; the results are summarized in column 2 in Table 2.2. The average decision time was higher under process accountability but with the predicted difference of only 0.4 seconds [-0.2, 1.1] per trial for a typical subject. In addition, the estimate is imprecise, meaning that the data are consistent with a wide range of effects. For example, the model assigns a probability of 0.08 to the possibility that process accountability *decreases* decision time.

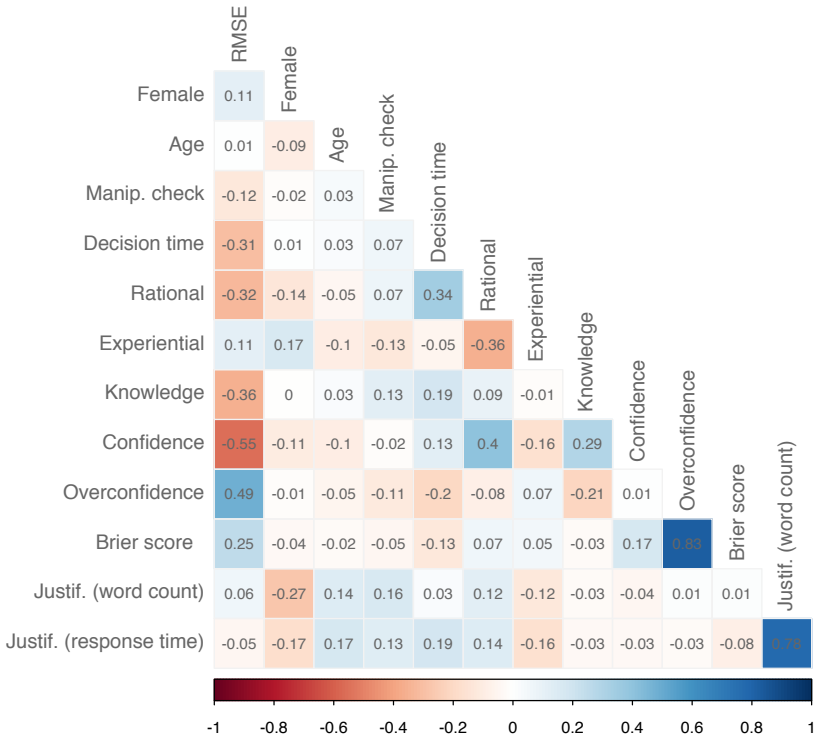


Figure 2.2: Correlations (Pearson's *r*) between the various measures. RMSE is root mean squared error (based on observed popularity), so a measure which positively predicts performance will correlate negatively with RMSE. All measures were computed using the full sample, except for Justification (word count) and Justification (sec); these measures are based on subjects in the process accountability conditions because only these subjects wrote justifications. See section *Appendix 2* for variable definitions. (Color version of this figure is available in the electronic version of this document.)

Table 2.2: Bayesian posterior means (standard deviations), full sample.

Parameter	Decision time	RMSE	Rational	Experiential	Knowledge	Brier score
Intercept	1.59 (0.03)***	1.86 (0.04)***	3.77 (0.05)***	2.84 (0.06)***	0.00 (0.06)	0.29 (0.01)***
Mixed Env.	0.04 (0.06)	-0.45 (0.07)***	-0.03 (0.10)	0.02 (0.12)	-0.41 (0.11)***	-0.08 (0.01)***
Process Acc.	0.08 (0.06)	-0.12 (0.07)*	0.15 (0.10)	-0.18 (0.12)	0.21 (0.11)*	0.02 (0.01)
M.E. x P.A.	-0.03 (0.12)	0.10 (0.15)	-0.07 (0.20)	0.36 (0.23)	-0.19 (0.23)	-0.00 (0.03)
Male	0.00 (0.06)	-0.16 (0.08)**	0.21 (0.11)**	-0.30 (0.12)**	-0.01 (0.12)	0.00 (0.02)

Prediction accuracy

We follow de Langhe et al. (2011) in using the *root mean squared error* as a measure of prediction accuracy, with lower values indicating better performance. The *squared error* part of the measure is based on *observed* popularity, which equals the true underlying popularity of the EasyPhone plus random error. The modeling results are summarized in column 3 of Table 2.2. There is a substantial main effect of environment, with higher accuracy in the mixed environment. There is also some evidence for a negative (meaning error-reducing) effect of process accountability. This result is directionally consistent with that of de Langhe et al. (2011), although we observe a substantially smaller effect size and considerable uncertainty. In particular, the results of de Langhe et al. (2011) imply a standardized effect size (Cohen's *d*) of about -0.64. In our study, the equivalent effect size (i.e. for the standard environment only) is -0.32. Using the model to quantify the uncertainty around this estimate, we obtain an interval of [-0.73, 0.05] and a probability of 0.93 that the population effect size in this study is less extreme than the one reported by de Langhe et al. (2011).

Note that because we are using the *observed* popularity to assess the accuracy of predictions, the measure is based only on the first part of the task (trials 1 - 110) and not the second part (final 32 trials), during which subjects did not receive any feedback. We therefore checked the robustness of the results by computing a second root mean squared error measure that encompasses all 142 trials. This measure is based on the *latent* popularity of EasyPhones, as determined by the equations reported in the section *Experimental manipulations*. The results based on this measure are very similar to the ones for our main measure.

Finally, in Figure 2.3 we plot the *moving* root mean squared error to obtain insight into how accuracy changed over time. We set the window to 22 trials and used the *latent* popularity when computing the root mean squared error so that we can observe accuracy in the last block of trials, when performance feedback ceased. The figure displays the averages for each experimental condition and the corresponding 95% uncertainty bands obtained using a boostrapping procedure. We observe a strong effect of environment and clear evidence of improvement in both environments until the final block, when the trend reverses. As expected given the sign of the interaction coefficient reported in column 3 of Table 2.2, the difference between the two accountability conditions is more pronounced in the standard environment—but only around the middle of the first part of the task.

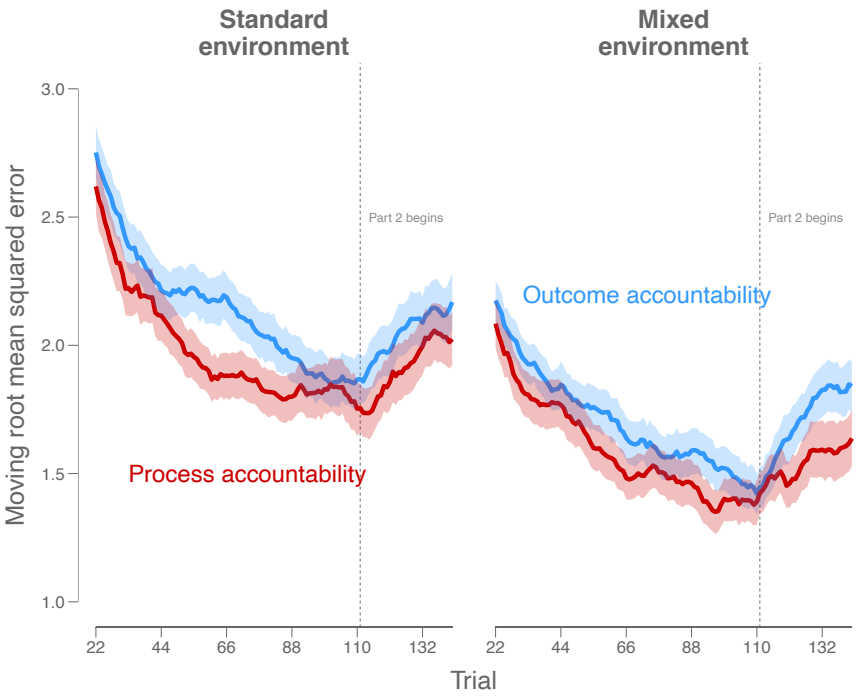


Figure 2.3: Accuracy over time. The measure depicted on the *y*-axis is the moving root mean squared error (computed by comparing subjects' judgments with *latent* popularity), so lower values better reflect better performance. The moving window was set to 22 trials. The colored lines denote condition averages, and the shaded areas display 95% uncertainty intervals obtained using a bootstrapping procedure (number of iterations: 1000). (Color version of this figure is available in the electronic version of this document.)

Confidence

Figure 2.4 compares the subjects' confidence in the accuracy of their predictions (defined as being within 0.5 of the actual popularity) with the actual accuracy rate. In the case of perfect calibration (dashed lines in the figure), subjects' confidence would perfectly track their accuracy. For example, a subject would be correct 75% of the time when their confidence was 75%. Overconfidence is indicated by the actual calibration curve (red and blue lines in the figure) being above the dashed line that denotes perfect calibration. Conversely, the actual curve being below the perfect calibration line indicates underconfidence. Our results replicate a well-known pattern in the literature of overconfidence for all but the most difficult trials (the trials for which the subjects are least confident). The figure also suggests that calibration is considerably better in the mixed environment compared to the standard environment. The difference between the two accountability conditions, however, appears small.

Next, we computed the *Brier score* (Brier, 1950), which is the mean squared difference between the probability assigned to an event and the actual outcome. In our case, the actual outcome equals one if the subject was accurate in a particular trial and zero otherwise. The Brier score tracks how much insight subjects have into their performance. The ideal score of zero would be achieved if a subject always expressed 100% confidence when they were in fact accurate, and 0% confidence when they were not. Siegel-Jacobs and Yates (1996) found lower Brier score among process-accountable subjects compared to outcome-accountable ones. Our analysis, reported in column 4 of Table 2.2, does not replicate this finding. The coefficent for process accountability is positive (meaning worse performance) and the model assigns a probability of only 0.13 to the hypothesis that process accountability improves the Brier score relative to outcome accountability. We do find, however, a strong main effect of environment. Mirroring the results for prediction accuracy, subjects in the mixed environment outperformed those in the standard environment.

Postexperimental questionnaire

We constructed an aggregate measure of knowledge by combining the two relevant sets of questions. First, we computed Pearson's correlation coefficient between actual cue weights and subject's estimates of cue importance[6] (*cue weight items*, see Procedure and materials). Second, we counted how often subjects correctly compared the two cue values within a single cue (*cue value items*). In the case of the *color* cue, for example, subjects who had indicated that blue EasyPhones were more popular than red ones received a score of one if that was indeed the case, and a score of zero if red EasyPhones were more popular *or* if there was no difference. Finally, an aggregate measure of explicit knowledge was constructed by first standardizing and then averaging these two preliminary measures. As expected, higher values of explicit knowledge predicted lower root mean squared error (see Figure 2.2). We regressed knowledge on the standard set of predictors; the results are summarized in column 5 of Table 2.2. First, we observe a strong main effect of environment; subjects in the mixed environment displayed worse knowledge than subjects in the standard environment. However, a follow-up

[6] Some subjects entered the same estimate for all four cues, yielding an undefined value of the correlation coefficient. We imputed the value of zero for these subjects.

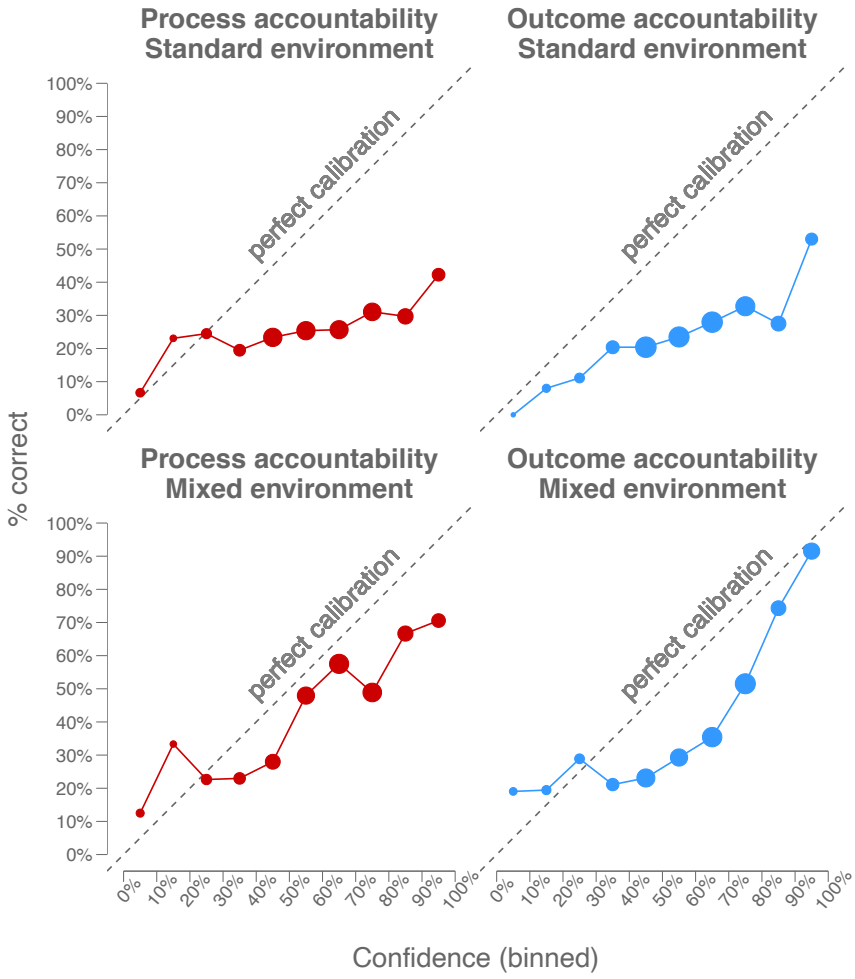


Figure 2.4: Calibration curves for the four experimental conditions. On the x-axis, subjects' probability judgments that their prediction was correct (i.e. within 0.5 of the truth) were binned into 0%-10%, 11%-20%, …, 90%-100%. On the y-axis, the actual percentage of correct predictions. The dashed line represents perfect calibration. The size of the points represents the number of probability judgments in the bin. (Color version of this figure is available in the electronic version of this document.)

analysis raised the possibility that this effect is illusory. Specifically, the analysis revealed that the effect is entirely due to the *cue value* questions. In particular, two of the cues in the mixed environment had a weight of zero, meaning that the correct answer to the *cue value* questions was that there was no difference between the two cue values for popularity. Returning to our *color* cue example, the correct answer would be that there was no difference in popularity between blue and red EasyPhones. Subjects did poorly on these questions, which could mean that they had difficulty learning that a cue did not matter at all, and would instead pick up illusory patterns. However, this claim rests on a measure that is based on the *latent* popularity whereas subjects only observed *observed* popularity. It is therefore possible that subjects picked up small but real differences originating from the noise component of the equation governing the popularity of EasyPhones. Unfortunately, the data were not recorded at a sufficiently granular level to allow us to test this possibility. The main effect of environment on the level of knowledge therefore cannot be unambiguously interpreted. In addition to the main effect of environment, the model also provides some evidence for a main effect of accountability. In particular, the level of knowledge was higher under process accountability than under outcome accountability.

Finally, we analyzed situation-specific thinking style. We averaged over the relevant trials to obtain two measures, one for *rational* thinking style and one for *experiential* thinking style (Novak and Hoffman, 2009). Figure 2.2 reveals that the two measures correlate negatively and that rational thinking style correlates more strongly with prediction accuracy. In particular, subjects with higher scores on this measures achieved lower error. Examination of the means in Table 2.1 shows that, in line with de Langhe et al. (2011), the means are higher under process accountability than under outcome accountability. As with prediction accuracy, however, the effect is weaker in magnitude. Whereas de Langhe et al. (2011) observe a standardized effect size (Cohen's *d*) of about 0.48, the equivalent effect size in the current study is 0.26 (model-based uncertainty interval: [-0.12, 0.68]) for the standard environment and 0.21 [-0.07, 0.52] overall. Comparing the raw means, we observe that in the process accountability conditions the current study replicates the results of de Langhe et al. (2011) down to the first decimal place. The difference is due to the outcome accountability conditions, where the subjects in the current study scored notably higher than their counterparts in de Langhe et al. (2011).

2.4 Discussion

The current study was motivated by the lack of independent replication studies in the literature on process and outcome accountability, coupled with the recognition of the essential role replications play in enabling a cumulative science. More specifically, our goal was to replicate and extend two key findings reported by de Langhe et al. (2011). We attempted to replicate the finding that *(i)* process accountability facilitates higher accuracy in elemental multiple-cue learning tasks compared to outcome accountability, and the finding that *(ii)* process accountability facilitates a more rational thinking style as measured by the scale of Novak and Hoffman (2009)—again relative to outcome accountability. We extended these findings in two ways. First, we collected additional measures, including decision time, confidence judgments, and a custom measure of explicit task knowledge. These measures have the potential

to provide additional insight into the mechanism underlying the performance differences between process and outcome accountability. Second, we tested the generalizibility of the findings by additionally investigating a variation of the task used by de Langhe et al. (2011). Whereas the original task (labeled *standard environment* in the current study) featured only cues that are actually informative and thus helped predict the criterion, the new version (*mixed environment*) featured two fully *uninformative* cues that did not help in prediction. We included this variation to test the robustness of the original findings and because some earlier accountability research (Tetlock and Boettger, 1989) as well as some of our related results (see section *Introduction*) indicated that the inclusion of such cues might moderate the positive effects of process accountability.

Did we replicate the two main findings of interest (de Langhe et al., 2011)? This question invites a binary answer and it is common to see the outcomes of replication studies described as either a "success" or a "failure." But in this case a more nuanced conclusion is called for. On the one hand, the results for both prediction accuracy and rational thinking are directionally consistent with those of de Langhe et al. (2011). On average, prediction accuracy was higher under process than under outcome accountability, and the mean level of rational thinking style was likewise higher under process accountability. On the other hand, the results of the current study are substantially weaker both in terms of the magnitude of the effects and in terms of statistical support. For both of the main results, our effect size is about half the size of that of de Langhe et al. (2011). Moreover, the considerable uncertainty around these estimates means that the data are compatible with a wide range of parameter values, including zero. In the framework proposed by LeBel et al. (2019), the results of this replication study would therefore be summarized as *no signal — consistent* because the 95% uncertainty intervals include both zero and the effect size observed by de Langhe et al. (2011).

The additional evidence obtained from measures of decision time and explicit task knowledge resonates with the main findings. We found that both mean decision time and mean level of knowledge were higher under process than under outcome accountability. In the language of dual-process theories of human cognition (Kahneman, 2011), these results are consistent with the idea that process accountability stimulates Type 2 processes more than outcome accountability. Again, however, the estimated effects of accountability type were small and noisy. This was especially true for decision time where the estimated difference was less than half a second more per trial.

In addition to collecting data from these additional measures, our study extended the results of de Langhe et al. (2011) by including a variation of the original task in which the weights of two of the cues were set to zero. In contrast to the effects of accountability type, we found clear evidence for strong effects of environment type. In particular, subjects were both more accurate and better calibrated in the mixed environment. We found no evidence, however, for any interaction between the type of environment and accountability type.

One final result deserves attention. Because we collected probability judgments in the second part of the task, we were able to additionally test whether process accountability improved calibration. This question was not investigated by de Langhe et al. (2011). But it was investigated by Siegel-Jacobs and Yates (1996), who found that process-accountable subjects were more accurate and better calibrated than outcome-accountable ones. We obtained no evidence for

this result in the current study. One major difference between the two studies needs to be highlighted, however. In the research of Siegel-Jacobs and Yates (1996) subjects were directly accountable for their probability judgments. In contrast, these judgments were secondary in the current task, as the primary task (for which the subjects were held accountable) was to accurately predict the criterion (the popularity of EasyPhones) and not to accurately formalize their confidence in the accuracy of their prediction (the probability task). Thus our results should not be interpreted as a failure to replicate the findings of Siegel-Jacobs and Yates (1996).

The current study has implications for future research building on the results of de Langhe et al. (2011). The small effect sizes observed in this study suggest that future studies should employ substantially larger sample sizes than what has been the norm so far in the accountability literature. Alternative approaches for increasing the precision of the estimates include the use of highly predictive covariates and the design of stronger manipulations. The substantial decrease in the observed effect sizes is interesting in light of the fact that the current study was a direct replication that also sampled from a very similar population of subjects as the original research (which was conducted at the same institution). At the same time, the current research features only a single study which puts a clear limit on how much we can learn. In our view, findings from the literature on process and outcome accountability would provide a very interesting target for a "Many Labs"-style collaborative investigation, in which a larger number of researchers conduct the same study using diverse samples (Klein et al., 2018). It would be interesting to benchmark the variation in effect sizes that would be observed against similar projects that have already been conducted, providing a test of our conjecture that accountability effects display considerable variability. Another highly informative approach would be to conduct a "metastudy" (Baribault et al., 2018). In this context, a metastudy would include a reasonable set of operationalizations of process and outcome accountability from which the exact operationalization for a given subject would be randomly sampled. Hierarchical Bayesian modeling could then be used to pool information across the resulting datasets and investigate the variability of the focal effect. A more modest recommendations for future research is to carefully document and share the research materials used in a study. As LeBel et al. (2019) note, a study cannot be properly scrutinized if it's not reported with sufficient transparency. The current project contributes to this by openly sharing all materials as well as data.

In conclusion, we hope that the present research stimulates future replication studies targeting key findings in the literature on process and outcome accountability.

2.5 Appendix 1

Table 2.3 provides an abstract representation of the stimuli used in the study. Each row corresponds to one EasyPhone and columns 3 - 6 define the corresponding EasyPhone in terms of the four cues (color, size, shape, and keyboard arrangement). Column 2 can be used to distinguish between the EasyPhones used in both parts of the study from those used only in the second part of the study (last 32 trials).

Table 2.3: Abstract representation of the stimuli.

ID	Used in:	Cue 1	Cue 2	Cue 3	Cue 4
1	both parts	0	0	0	0
2	both parts	1	0	0	0
3	both parts	0	1	0	0
4	both parts	1	1	0	0
5	both parts	0	0	1	0
6	part 2	1	0	1	0
7	both parts	0	1	1	0
8	both parts	1	1	1	0
9	both parts	0	0	0	1
10	part 2	1	0	0	1
11	both parts	0	1	0	1
12	both parts	1	1	0	1
13	both parts	0	0	1	1
14	part 2	1	0	1	1
15	both parts	0	1	1	1
16	part 2	1	1	1	1

2.6 Appendix 2

Table 2.3: Variable definitions.

Label	Definition
Manipulation check	The average of two outcome accountability-related self-report items minus the average of the two process accountability-related self-report items. The items were administered in the postexperimental questionnaire and used a 5-point Likert response scale. See section *Procedure and materials* for the exact items.
Decision time (seconds)	The average time that elapsed between stimulus presentation and the submission of the prediction; outliers were excluded beforehand. See section *Predecisional behavior*.
RMSE	Root mean square error of the predictions, with the error computed based on *observed* popularity (trials 1 - 110). See section *Experimental manipulations*.
Rational thinking	Average of the (slightly adapted) ten items of the rational subscale of the situation-specific thinking style questionnaire (Novak and Hoffman, 2009). The response scale was a 5-point Likert scale. The exact items are available in the project repository (see section *Additional information*).
Experiential thinking	Average of the (slightly adapted) ten items of the experiential subscale of the situation-specific thinking style questionnaire (Novak and Hoffman, 2009). The response scale was a 5-point Likert scale. The exact items are available in the project repository (see section *Additional information*).
Knowledge	Score of explicit task knowledge based on a quiz included in the postexperimental questionnaire. In one set of questions, subject estimated the *cues* (e.g. how important the color cue was). In another set of questions, they estimated cue *values* (e.g. were blue phones more popular than red ones, vice versa, or no difference). The exact items are available in the project repository (see section *Additional information*). For the exact calculation of the knowledge measure, see section *Postexperimental questionnaire*.
Confidence	Average confidence in the prediction being within 0.5 of actual popularity. Based on the 32 trials of the second part of the main task.
Overconfidence	Confidence (see above) minus the actual percentage of the 32 trials that were within 0.5 of actual popularity.

Label	Definition
Brier score	Average squared difference between the confidence in the prediction and the truth (the latter is 1 if the prediction was within 0.5 of the actual popularity, 0 otherwise) a given trial. Based on the 32 trials of the second part of the main task.
Justification (word count)	Number of words written in response to the *justify-your-approach* question.
Justification (seconds)	Time (in seconds) taken to submit the response to the *justify-your-approach* question

2.7 Appendix 3

As noted in the section *Preliminary analyses*, there were nine subjects (six in the process accountability condition, three in the outcome accountability condition) for whom the manipulation check score was directionally inconsistent with their accountability condition. Table 2.4 replicates Table 2.2, except that these nine subjects were excluded from the sample before fitting the models.

In addition to the nine subjects mentioned above, there were 29 subjects (16 in the process accountability condition, 13 in the outcome accountability condition) for whom the manipulation check score was a neutral zero. If we *additionally* exclude these subjects before fitting the models, we obtain the results summarized in Table 2.5.

Table 2.4: Bayesian posterior means (standard deviations), reduced sample 1.

Parameter	Decision time	RMSE	Rational	Experiential	Knowledge	Brier score
Intercept	1.59 (0.03)***	1.85 (0.04)***	3.79 (0.05)***	2.82 (0.06)***	0.01 (0.06)	0.29 (0.01)***
Mixed Env.	0.04 (0.06)	-0.43 (0.07)***	-0.02 (0.10)	0.01 (0.12)	-0.44 (0.11)***	-0.09 (0.02)***
Process Acc.	0.06 (0.06)	-0.12 (0.07)*	0.17 (0.10)*	-0.17 (0.12)	0.23 (0.12)**	0.02 (0.02)
M.E. x P.A.	-0.02 (0.12)	0.12 (0.15)	-0.14 (0.19)	0.38 (0.23)	-0.17 (0.23)	-0.01 (0.03)
Male	-0.01 (0.06)	-0.15 (0.08)*	0.18 (0.10)*	-0.27 (0.12)**	-0.03 (0.12)	0.00 (0.02)

Table 2.5: Bayesian posterior means (standard deviations), reduced sample 2.

Parameter	Decision time	RMSE	Rational	Experiential	Knowledge	Brier score
Intercept	1.58 (0.03)***	1.87 (0.04)***	3.74 (0.06)***	2.81 (0.07)***	-0.00 (0.07)	0.29 (0.01)***
Mixed Env.	0.07 (0.06)	-0.40 (0.08)***	-0.03 (0.10)	0.05 (0.13)	-0.50 (0.13)***	-0.07 (0.02)***
Process Acc.	0.09 (0.06)	-0.11 (0.08)	0.10 (0.10)	-0.28 (0.13)**	0.29 (0.13)**	0.01 (0.02)
M.E. x P.A.	0.03 (0.13)	0.08 (0.16)	-0.12 (0.21)	0.41 (0.25)	-0.08 (0.27)	-0.01 (0.03)
Male	-0.03 (0.07)	-0.14 (0.09)	0.10 (0.11)	-0.20 (0.13)	0.00 (0.14)	-0.01 (0.02)

Chapter 3

On insight problem solving under process and outcome accountability

There ain't no entertainment
And the judgements are severe
— Leonard Cohen describes environments that stifle creativity

Summary:

The effects of accountability on judgment and decision making have received a lot of attention in the past few decades. This research suggests that holding individuals *process accountable* typically results in better performance than holding them *outcome accountable*. However, the effects of different types of accountability on *creative problem solving* remain unexplored. In this chapter, we address this gap by studying the effects of process and outcome accountability on performance on *insight problems*—problems which tend to be challenging because they invoke a misleading cognitive representation or invite incorrect solution attempts. Drawing on the literature on accountability and insight, we test the prediction that in the domain of insight problem solving, performance will be better under *outcome* accountability. We report the results of three (plus two) experiments using two commonly used insight tasks, the matchstick arithmetic task and the remote associates task. Although we obtain preliminary evidence that the prediction may hold for the remote associates task, altogether, the results are mixed and highlight the inadequacy of the proposed theoretical account and the fickle nature of accountability effects in this domain. We summarize the main lessons of the current research and close by emphasizing the role of theory development, independent replication studies, and open research practices in driving the future progress in this research program.

Acknowledgments:

Experiment 1 was designed together with Franziska Grotheer, who also collected the first wave of Experiment 1 data and used it in her Master thesis. We thank her for her contribution. We also thank Philip Eskenazi for sharing the data and the stimuli from his experiments, and Evelien Reusen for her help in recruiting the subjects for Experiment 2.

3.1 Introduction

Talking about research and development to executives at Disney[1], the pioneering computer scientist Alan Kay once described some of the ways to "kill the geese that lay the golden eggs." These include setting up deadlines and quotas, asking the "geese" to "justify their diet and day to day [*sic*] processes," and demanding that they explain "just how they will make the eggs that will be laid." For designers of management control systems, Kay's observations highlight an uncomfortable tension between the desire for valuable creative output and the demand for accountability. Traditional approaches that are effective at facilitating run-of-the-mill productivity may not be equally effective at stimulating fruitful insights that would solve challenging problems faced by businesses and other organizations.

We define accountability as an expectation that (some aspect of) one's performance will be evaluated by an audience (such as one's manager) and that the evaluation will result in positive or negative consequences for the accountable individual (Siegel-Jacobs and Yates, 1996). We start from the basic premise that some degree of accountability is inescapable in organizational life. This premise shifts the question from *if* accountability should be implemented to *how* (Tetlock et al., 2013). In response to this question, the literature in psychology and management highlights a key distinction between *process* and *outcome* accountability (Patil et al., 2014). Whereas process-accountable individuals are expected to justify *how* they approached a given problem and evaluated on the reasonableness of their approach, outcome-accountable individuals are held accountable for the eventual outcomes of their actions (Simonson and Staw, 1992).

In the last two decades or so, a number of laboratory studies investigated the effects of process and outcome accountability on accuracy as well as other outcomes of interest to management scholars, such as self-serving behavior (Pitesa and Thau, 2013), negotiation tactics (Chang et al., 2013), and professional skepticism during an audit (Kim and Trotman, 2015). These studies span multiple academic disciplines, including psychology (Siegel-Jacobs and Yates, 1996), personnel management (Brtek and Motowidlo, 2002), and management accounting (Dalla Via et al., 2018). Their results suggest that process accountability is typically more effective at stimulating accuracy and other desirable outcomes than outcome accountability (Patil et al., 2014) and also provide some evidence regarding the underlying cognitive mechanism that mediates these findings (de Langhe et al., 2011).

Although these studies span a number of domains of clear theoretical and practical interest, the domain of creative performance has been relatively neglected. The only exception we are aware of is the study by Häusser et al. (2017), who contrasted the effects of process and outcome accountability (separately and in combination) on idea generation and selection. Although the mean creativity ratings of the generated solutions were similar across the conditions, subjects in the no accountability condition produced more unique ideas than subjects under (any kind of) accountability—providing some corroboration to Alan Kay's conjecture summarized in the first paragraph.

Our research differs from the work of Häusser et al. (2017) as we consider a different aspect of

[1] From: http://worrydream.com/2017-12-30-alan/, accessed on April 20, 2020

creative problem solving. In general, problems that call for creative solutions can be of two types. Some problems are relatively open-ended, such that many different (perhaps radically so) solutions could in principle fulfill the goals of the individual or the organization. For example, many different advertising campaigns could successfully support the marketing of a new product. Problems of this type were the focus of Häusser et al. (2017). Other challenging problems impose tight constraints on the solution space. In the extreme, only one solution will do. In this chapter, we investigate a class of problems belonging to this second category, called *insight* problems in the psychology literature. Although we directly focus on these problems only, some aspects of insight problem solving may generalize to more open-ended creative problem solving too.

Our interest in insight problems is a function of two factors. First, insight problems can serve as a model for important discoveries in business and science. Therefore, learning which factors support or hinder successful performance on these problems in laboratory conditions has the potential to inform the design of management systems in real world settings, where it may be very challenging to assess their effects in this domain directly (compared to simpler forms of productivity). Second, the literature on insight and the results of Eskenazi (2015) raise an interesting possibility that the commonly observed superiority of process over outcome accountability will not extend to the domain of insight problems.

In the remainder of this *Introduction*, we first describe insight problems and summarize the most relevant theoretical ideas and experimental results of the literature on this topic. Next, we provide an overview of the literature on process and outcome accountability and highlight the dominant explanation for the common finding of higher performance under process accountability than under outcome accountability. Finally, we tie these literatures together and discuss what they jointly suggest about insight problem solving under process and outcome accountability.

Insight problem solving

Failure in problem solving can have different causes. For example, the solution to a problem could prove elusive because the individual simply lacks the knowledge or ability that is required to solve the problem. Alternatively, the source of difficulty could reside in the way in which the individual thinks about the problem. People are sometimes stuck because they frame the problem in a way that is unnecessarily constraining or otherwise unhelpful. If an inappropriate *cognitive representation* of the problem is adopted, people may fail to produce a solution even when they in fact possess the required knowledge, ability, and motivation (Knoblich et al., 1999).

Problems that tend to elicit unhelpful cognitive representations are often called *insight problems*. That is because, when these problems *do* get solved, a particular sequence of cognitive activity is typically reported. It includes initial but unsuccessful attempts, a state of *impasse* during which an individual is stuck and does not know how to move forward, and finally the sudden and unanticipated grasping of the solution (Chu and MacGregor, 2011). Many discoveries and inventions in domains ranging from science to art fit this pattern, which may explain why the study of insight has become a lively interdisciplinary research area in recent decades.

How do individuals solving insight problems eventually change the cognitive representation of the problem from the initial, unproductive one to a representation that facilitates the finding of the solution? The answer to this question is the subject of an ongoing theoretical debate but according to the influential *special-process* view of insight, the change in representation that enables insight is brought on by processes that are implicit, automatic, and unconscious in nature, or *Type 1* in the terminology of modern dual-process theories of human cognition (Knoblich et al., 1999; Bowden et al., 2005). According to dual-process theories, Type 1 processes differ fundamentally from processes of *Type 2* which are conscious, deliberate, effortful, and mediated by working memory (Kahneman, 2011). The special-process view of insight suggests that an increase in Type 2 (controlled, deliberate) processing may not be as beneficial in insight problems as it is in other kinds of problems. The stronger version of this hypothesis states that an increase in Type 2 can even prove *detrimental* to the extent that it interferes with Type 1 processing that would otherwise lead to a representational change.

This intriguing hypothesis has been studied in a number of ways. Some studies measure working memory capacity and correlate it with performance on insight problems (Fleck, 2008; Chein et al., 2010; Chein and Weisberg, 2014; DeCaro et al., 2016; Chuderski and Jastrzębski, 2018). The results of these studies are mixed and subject of an on-going debate (Chuderski and Jastrzębski, 2017; DeCaro et al., 2017). But on the whole, the results seem more consistent with the idea that working memory capacity correlates *positively* with performance on insight problems, though DeCaro et al. (2017) argue for the existence of important moderating factors. In any case, the correlational nature of this evidence presents some difficulties for interpretation. We therefore focus on studies that employed an experimental approach to investigate the role of Type 1 and Type 2 processes in insight problem solving.

The basic logic of these experiments is to manipulate the extent to which individuals rely on Type 1 and Type 2 processes and observe the downstream consequences on performance. Some of these studies attempt to *increase* the relative extent of Type 2 processing by asking subjects to verbalize their thinking during problem solving (Schooler et al., 1993; Ball and Stevens, 2009; Gilhooly et al., 2010; Ball et al., 2015). Other studies attempt to *decrease* the relative extent of Type 2 processing by employing a concurrent secondary task while subjects solve insight problems (Lavric et al., 2000; Ball and Stevens, 2009; Jarosz et al., 2012; Ball et al., 2015; DeCaro and Van Stockum, 2017). Alternative approaches that follow the same basic idea include using alcohol intoxication (Jarosz et al., 2012) or relying on the ego depletion paradigm (DeCaro and Van Stockum, 2017). Finally, Zedelius and Schooler (2015) attempted to both increase (in one condition) and decrease (in the other) the relative extent of both Type 1 and Type 2 processing by instructing subjects to either engage in an active, "analytical" approach to problem solving or in a more passive, "intuitive" approach.

The majority of these experiments are consistent with the hypothesis highlighted above. That is, a relative increase (decrease) in Type 2 (Type 1) processing was associated with either a decrease in performance on insight problems (stronger version of the hypothesis) or at least with a lack of an increase in performance (weaker version of the hypothesis). Of course, studies in this second category offer only limited evidence, as the lack of an increase can indicate nothing more than insufficient statistical power. In addition, the results of at least one study (Ball and Stevens, 2009) directly contradict the hypothesis, as verbalization (meant to increase reliance on Type 2 processing) was linked to improved performance whereas auditory suppression

(meant to decrease reliance on Type 2 processing) was linked with decreased performance. Therefore, the results in this stream of literature are overall not conclusive. But, at the very least, they do indicate that in some circumstances an increase in deliberate, controlled cognitive processing may not pay off in terms of improved performance on (some) insight problems.

Process and outcome accountability

Under pure *process* accountability, the evaluative focus is on the individual's judgment strategy, problem-solving approach, or reasons for making a specific decision. Under pure *outcome* accountability, these considerations are ignored and the only thing that matters are the relevant consequences for the organization, such as the profitability of one's investment strategy. Clearly, there are many ways in which each of these two types of accountability could be implemented. For example, a management control system based on the idea of process accountability could be linked with precise guidelines that the individuals are supposed to follow. The effectiveness of such a system would in large part depend on the appropriateness of those guidelines for the domain at hand (Patil et al., 2017). The focus in the existing literature on accountability, however, has been on process accountability coupled with a lack of clear guidelines for action. It is this kind of process accountability that has been associated with superior performance compared to outcome accountability, and it is this kind of process accountability that we investigate in this chapter.

Why does this "open" form of process accountability appear to be more effective in facilitating accurate performance than the more direct approach of outcome accountability? In one of the early demonstrations of this effect (in the domain of probability judgment), Siegel-Jacobs and Yates (1996) suggested that process-accountable individuals on average process a larger amount of information than outcome-accountable ones, although their own process-tracing data (average number of cues inspected per judgment) failed to provide evidence for this conjecture. However, a survey conducted among organizational buyers did reveal some correlational evidence for a link between process accountability and information collection and analysis (Doney and Armstrong, 1996). Later experimental studies suggested that although the absolute amount of information considered may not reliably differ under process and outcome accountability, the *way* this information is processed might. Relying on real-time observation of behavior by a confederate, Brtek and Motowidlo (2002) reported that process-accountable subjects appeared more *attentive* when analyzing video-taped interviews for the purpose of providing predictions on how the interviewed candidates would perform on the job. More recently, Häusser et al. (2017) reported that time spent on the task was higher under process accountability and Dalla Via et al. (2018) reported a larger number of saccades under process accountability (interpreted as *search effort* by the authors).

These findings link the accountability literature to dual-process accounts of human cognition introduced in the previous section, as an increase in time spent on the task and number of saccades/fixations are markers for increased engagement of Type 2 processing (e.g. Horstmann et al., 2009; Dickert et al., 2012). But the clearest evidence yet comes from a series of studies by de Langhe et al. (2011). These authors found that process-accountable subjects were more accurate than outcome-accountable ones in an elemental multiple-cue judgment task (meaning

that the relationship between predictive cues and the criterion was linear). Crucially, this difference was *moderated* by personality differences in the tendency to engage in analytical thinking and *mediated* by situation-specific rational thinking style, suggesting that — at least in their task — the superiority of process accountability can be explained by the more deliberate and systematic cognitive processing of Type 2 on the part of process-accountable individuals. Additional results consistent with this hypothesis come from investigations that focused specifically on process accountability (e.g. Scholten et al., 2007).

Accountability focus and insight problem solving

The empirical work reported in the previous section suggests that process accountability is typically more effective in facilitating the deliberate, controlled cognitive processing of Type 2 than outcome accountability (de Langhe et al., 2011). In light of this conjecture, the fact that performance is often higher under process accountability (Patil et al., 2014) is perhaps not surprising. That is because a large body of evidence in psychology identifies insufficient engagement of Type 2 processing (in response to the automatic outputs supplied by Type 1 processing) as a major culprit behind many failures of human thinking (Kahneman, 2011).

But although an increase in Type 2 processing leads to improved performance in a variety of situations, this is not universally the case. Drawing on a diverse body of work in cognitive science, psychology, and neuroscience, Amer et al. (2016) show that some forms of learning and problem solving can in fact be *impaired* by an increase in the controlled processing of Type 2. For example, they argue that "reduced cognitive control in older adults may boost creativity and their ability to solve insight problems." This claim is consistent with (most of the) experimental evidence reviewed above, as older age is associated with a decline in controlled processing. Overall, this work raises the possibility that the relative advantage of process over outcome accountability may be diminished or even reversed in the case of insight problem solving.

In the accountability literature, evidence consistent with this hypothesis was reported by Eskenazi (2015, Chapter 3). Across three studies, the average performance on the remote associates task was consistently lower in the process accountability condition than in the no accountability condition. Because the remote associates problems are challenging problems that are frequently solved with insight (Bowden and Jung-Beeman, 2003), this result supports the above hypothesis if we assume that outcome accountability would not have had such a performance-decreasing effect. The work reported in this chapter expands on the results of Eskenazi (2015) in three main ways.

First, we are the first to directly contrast process and outcome accountability in the domain of insight problems. We focus on this comparison because the typical question in organizational life is not *whether* accountability should be implemented but rather *how* (Tetlock et al., 2013). Another reason to directly compare process and outcome accountability has to do with the potential of accountability to induce stress (Siegel-Jacobs and Yates, 1996; Häusser et al., 2017) and the fact that stress has been linked to impaired performance on insight problems (e.g. Martindale and Greenough, 1973; Alexander et al., 2007). These observations raise

the possibility that process and outcome accountability would have a similarly deleterious performance effect to the extent that both would result in stress (Häusser et al., 2017).

Second, we attempt to replicate the finding of increased Type 2 processing under process accountability compared to outcome accountability (de Langhe et al., 2011) and link it to performance on insight problems. That is, we attempt to provide direct mediational evidence that the assumed differences among accountability conditions in insight problem solving can be explained by the differences in the degree of Type 1 and Type 2 processing among these conditions.

Third, we rely on two different types of insight problems to investigate our research question. Unfortunately, there are no problems that can elicit the insight sequence with perfect reliability. Indeed, the inconsistent results reported in the literature on insight problem solving could in part stem from the use of different kinds of problems in different studies. In the absence of a "gold standard," we tested two different types of problems in the experiments reported in this chapter, namely the *matchstick arithmetic* problems (Knoblich et al., 1999) and the *remote associates* problems (Mednick, 1962). These are among the most commonly used problems in research on insight and have certain methodological advantages over other candidate problems (Chu and MacGregor, 2011). At the same time, there are clear differences between these two types of problems. For example, matchstick arithmetic problems are quasi-mathematical whereas remote associates are verbal. Observing the same results using both of these problem types would enhance our confidence in their generality. Conversely, observing that the results differ would suggest the existence of moderators deserving of further study.

In the rest of this chapter, we report the results of three main and two supplementary experiments targeting our research questions before concluding with a general discussion in section 3.5.

3.2 Experiment 1

To the best of our knowledge, Experiment 1 is the first direct comparison of process and outcome accountability in the domain of insight problems. In addition to process accountability and outcome accountability, the experiment features two other experimental conditions. First, the control condition includes no explicit accountability manipulation and therefore helps us to better understand the nature of any differences between process- and outcome accountability conditions that we might observe. Second, the mixed accountability conditions models the common organizational setting that includes elements of both process and outcome accountability.

The task we have chosen for this first test features *matchstick arithmetic problems* (Knoblich et al., 1999). These are false arithmetic statements that are represented using Roman numerals composed of matchsticks. The goal is to transform a false statement into a mathematically correct one by moving a single matchstick. On the face of it, matchstick arithmetic problems ought to be easy because only a single move is needed to solve a given problem. In spite of this, subjects often encounter *impasse* and fail to solve (some of) the problems in the allocated time (Knoblich et al., 2001).

Knoblich et al. (1999) explain the difficulty of matchstick arithmetic problems by highlighting the role of common arithmetic and algebraic knowledge that is automatically activated upon encountering problems of this kind. This knowledge includes constraints on the possible operations and end states that prove counterproductive in this particular context. In addition, a solution move is less likely to be considered if it involves breaking an existing element into two subelements that are not meaningful on their own. For example, subjects are more likely to think of removing a matchstick from the Roman numeral III (an operation that produces two meaningful subelements, I and II) than from the Roman numeral X (an operation that produces two tilted matchsticks which are not meaningful on their own). On the basis of these considerations, Knoblich et al. (1999) were able to predict the relative difficulty of different types of matchstick problems. In particular—and using the labels reproduced in Table 3.1—the two distinct theorized orderings are $C > B > A$ and $D > A$, where the problem type on the left-hand side of the inequality is predicted to be the more difficult one. While our task features matchstick arithmetic problems of varying levels of difficulty, our primary interest is in the more difficult problems as these are the ones most likely to require restructuring and insight.

Methods

Subjects. A total of 163 subjects (recruited from a paid subject pool) participated in the study. The average age was 22.7 years (SD = 2.4), 102 subjects were female and 61 male. Participation in the study took about 45 minutes and subjects were compensated with 7.5€ for their time. The data for the first 100 subjects were collected in May and June 2017. To reach the planned sample size of 160, we conducted additional sessions between January and March 2018.

Experimental design. Subjects were randomly assigned to either a *no accountability* ($n = 41$), *process accountability* ($n = 40$), *outcome accountability* ($n = 41$), or *mixed accountability* condition ($n = 41$). Because the maximum bonus amount (see *Experimental manipulations* below) was the same in all three accountability conditions, this is not a pure 2 (process accountability: yes/no) × 2 (outcome accountability: yes/no) design.

Experimental manipulations. Right before the main task, subjects in the three accountability conditions read the text that served as a manipulation of accountability. They were informed that they were going to be evaluated by a team of judges and that they could earn a bonus of up to 7.5€ in addition to the basic payment. Subjects in the process accountability condition were told that this bonus was tied to how clear and detailed a description of their problem-solving strategy they produced and not to the number of problems solved correctly. In contrast, subjects in the outcome accountability condition were told that their bonus depended on the number of problems solved correctly and not on the way they solved them. These manipulations are similar to other manipulations of process and outcome accountability in the recent literature (de Langhe et al., 2011; Patil et al., 2017). Finally, subjects in the mixed accountability conditions were told that the bonus depended both on how many problems they solvde and on how clear and detailed a description of their problem-solving strategy they produced (each contributed 50%). Subjects in the control condition received no accountability-related instructions and proceeded directly to the matchstick arithmetic task.

Procedure and materials. After being welcomed and given the informed consent statement,

subjects proceeded to individual cubicles. From this point onward, the procedure was fully computerized. After entering demographic information and completing a 6-item measure of state anxiety (Marteau and Bekker, 1992), subjects practiced recognition of Roman numerals in three blocks of 15 trials (in each trial, the subject saw a Roman numeral, entered its Arabic counterpart, and received accuracy feedback). Following the training, subjects were given instructions for the matchstick arithmetic task, practiced selecting and moving the matchsticks using mouse and keyboard, and completed a 6-item quiz concerning task rules. The subjects in the accountability conditions than received the text that served as a manipulation of accountability, after which the main task began. This main task consisted of 12 problems taken from Knoblich et al. (1999) and listed in Table 3.1. The trials were divided into two blocks, with the order of trials within each block determined randomly for each subject. Each trial consisted of a matchstick equation and a timer counting down from 120 seconds. Once the subject pressed ENTER or after 120 seconds elapsed, subjects were prompted to enter the solution. The time to enter the solution was limited (10 seconds to select the matchstick and then 25 seconds to move it to the correct position) and no feedback was given afterwards. The trials automatically followed each other, except that there was a break after the first block. Upon completion of the matchstick arithmetic task, subjects in all conditions entered a description of their problem-solving strategy. Finally, subjects completed the postexperimental questionnaire. The items included the manipulation check and the payment expectation item (these two items were answered by subjects in accountability conditions only), an item about familiarity with matchstick arithmetic problems, a second administration of the 6-item state anxiety measure, three items concerning how much the subject thought about his or her task approach during the task (*process thinking*) and one item about time pressure. The study was programmed in Python using the PsychoPy package (Peirce, 2007).

Table 3.1: The twelve matchstick problems used in Experiment 1.

	ID	Problem	Solution
1	A1	VI = VII + I	VII = VI + I
2	A2	IV = III + III	VI = III + III
3	B1	I = II + II	I = III - II
4	B2	IV = III - I	IV - III = I
5	C1	III = III + III	III = III = III
6	D1	XI = III + III	VI = III + III
7	A3	II = III + I	III = II + I
8	A4	IX = VIII + III	XI = VIII + III
9	B3	III = V + III	III = VI - III
10	B4	V = III - II	V - III = II
11	C2	IV = IV + IV	IV = IV = IV
12	D2	VI = VIII + III	XI = VIII + III

Note: The problems were taken from Knoblich et al. (1999). Problems 1 - 6 appeared in Block 1, the rest in Block 2. The letter in the problem ID corresponds to problem type using the classification of Knoblich et al. (1999).

Scoring. Correctness of subjects' solutions was evaluated by examining the screenshots (covertly) taken during the experiment. Overall, in 57.2% of the trials the subject's solution matched the solution reported in Table 3.1. In addition, we identified seven other solution types. Four of these were explicitly forbidden in the instructions. They include obtaining the solution by *(i)* removing the matchstick off the screen (0.5% of all trials), *(ii)* using a non-standard Roman numeral (2.4%), *(iii)* using a not-equals sign (0.4%), and *(iv)* interpreting a clearly "tilted" matchstick as being non-tilted (i.e. vertical or horizontal) (0.9%). The remaining three solution types were not explicitly forbidden in the instructions. These include solutions obtained by *(i)* placing a matchstick directly on top of another matchstick (functionally equivalent to moving a matchstick off the screen) (0.3%), *(ii)* interpreting a "non-tilted" (i.e. vertical or horizontal) matchstick as being tilted (6.3%), and *(iii)* leaving the equation with a "hanging" operator (e.g. III = III +) (1.7%). For our primary analysis we accept as correct only those answers that match the solution reported in Table 3.1. But we also analyze the data using an alternative scoring rule that additionally accepts the three solution types not explicitly forbidden in the instructions.

Data analysis. We used Bayesian estimation to analyze the data. In particular, we fit a mixed-effects logistic regression model using the following specification:

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(\theta_{ij}) \\ \text{logit}(\theta_{ij}) &= \alpha_{\text{o}} + \alpha_{\text{subject[i]}} + \alpha_{\text{problem[i]}} + \\ &\quad (\beta_{\text{PA}} + \beta_{\text{PA, problem}[j]})\text{PA}_i \\ &\quad (\beta_{\text{OA}} + \beta_{\text{OA, problem}[j]})\text{OA}_i \\ &\quad (\beta_{\text{MA}} + \beta_{\text{MA, problem}[j]})\text{MA}_i, \end{aligned} \tag{3.1}$$

where y_{ij} equals one if subject i correctly solved problem j and zero otherwise, PA, OA, and MA are dummy variables corresponding to the process accountability, outcome accountability, and mixed accountability conditionss, α_{o} is the grand intercept, $\alpha_{subject}$ is the varying intercept for subjects, $\alpha_{problem}$ is the varying intercept for problems, and the final three lines correspond to the varying effects of the three accountability conditions (broken down into average effects and problem-specific adjustments).

Because the model is a Bayesian one, its parameters require a prior distribution. We used the results of a pilot study and domain knowledge to formulate weakly informative priors. In particular, we placed a normal prior with a mean of 0.5 and a standard deviation of 1 (i.e. Normal(0.5, 1)) on the grand intercept, reflecting the assumption that a typical subject placed in the control condition will correctly solve about 62% (95% uncertainty interval: [18%, 92%]) of the problems. For the average effects of the accountability conditions, we used a Normal(0, 0.5) prior. If someone's probability of solving a given matchstick problem equals 50%, this prior suggests that any of the three accountability manipulations may increase or decrease this probability by up to 23 percentage points, and that it is unlikely ($p < 0.05$) that the magnitude of the effect would be larger than that. The priors for the standard deviations were half-normal (truncated at zero) with a mean of 0 and standard deviation of 1 (subjects) or 1.5 (problems). Finally, an LKJ prior (Stan development team, 2017) with η set to 2 was used for the correlation matrix for the varying effects.

We fit the model using the R package `brms` (v2.4.0, Bürkner, 2017), which serves as an interface to the `Stan` language (Gelman et al., 2015).

For the benefit of our frequentist readers, we also attempted to fit an equivalent model using maximum likelihood estimation with the R package `lme4` (v1.1-21, Bates et al., 2015). Because the model did not converge, however, we fit a simpler model by dropping the varying slopes. The results are reported in section *Appendix 2*.

Additional information. Additional information, including study materials and data, is available in the project's online repository (`https://osf.io/mcxqk/`).

Results

Preliminary analyses. To investigate subjects' ability and motivation to solve the matchstick arithmetic task, we examined accuracy in the training task, number of correct answers on the comprehension quiz, and number of trials in which the subject terminated the trial before the time was up but subsequently did *not* enter any solution. This last measure can serve as a proxy for low solution-oriented motivation. This is more likely to be the case, however, if the trial was terminated substantially early as opposed to a second or two before the time is up. We therefore chose a more conservative criterion according to which the trial had to be terminated at least 10 seconds before the time was up in order to count as *early termination*.[2]

Summary statistics on all three measures are available in Table 3.2. Most subjects did well on the Roman numerals training and on the comprehension quiz. For example, 97.5% of subjects made less than five errors in the 45 trials of training and 79.1% of subjects made no more than one error on the comprehension quiz. For early terminations, however, an interesting pattern emerged. Subjects in the process accountability condition had more such trials than subjects in the other three conditions. To investigate this pattern, we fit a modified version of Model 3.1 (see page 52), changing the dependent variable (early termination instead of accuracy) and the priors.[3] The results suggested that, relative to the other three conditions, process accountability increased the probability that a trial will be terminated early (the associated posterior probabilities being 0.98/0.89/0.91 for the comparison with the control/outcome accountability/mixed accountability conditions). We return to this issue in the *Discussion*.

Two additional analyses served as manipulation checks. First, subjects in the accountability conditions were directly asked what their bonus depended on (for example, for outcome-accountable subjects the correct answer would be "number of problems solved correctly"). Overall, 86.1% of subjects answered correctly. In the main text, we report the results for the full sample because excluding the 17 subjects who answered incorrectly would introduce bias to our estimates (favoring the accountability conditions relative to the control condition)—but see section *Appendix 2*.

Second, we compared the conditions with respect to the average effort expended in justifying one's approach to the task, as proxied by writing time and word count. Financially, the justification stage was important for process-accountable subjects (as the entirety of their bonus

[2] This only decreased the number of early terminations from 135 to 126.

[3] We used the default priors in `brms` (v2.4.0, Bürkner, 2017).

depended on the evaluation of their justification), followed by mixed accountability subjects (for whom it was 50%). We would thus expect effort to be highest in the process accountability condition, followed by the mixed acountability condition, followed by the remaining two conditions. The results are in line with this prediction for both the number of seconds spent writing the justification and the number of words written.[4]

Table 3.2: Condition means (standard deviations) for various measures used in Experiment 1, separately for each experimental condition (see section *Appendix 1* for definitions).

	Control	Process	Outcome	Mixed
Gender (% female)	53.7	65.0	68.3	63.4
Age (years)	23.3 (1.9)	22.2 (2.6)	22.4 (2.5)	22.9 (2.7)
Training	43.9 (1.4)	43.9 (2.0)	43.9 (1.3)	44.0 (1.3)
Comprehension quiz	5.1 (0.9)	5.1 (1.0)	5.0 (0.9)	5.0 (0.8)
Manipulation check (% pass)	/	82.5	90.2	85.4
Response time (seconds)	57.4 (13.0)	55.4 (11.8)	56.4 (13.2)	51.9 (10.7)
Early terminations	0.5 (1.6)	1.2 (1.8)	0.7 (1.3)	0.7 (1.6)
Total score	7.1 (2.2)	6.3 (2.4)	7.1 (2.5)	6.9 (2.1)
Total score, alternative scoring	8.2 (2.0)	7.2 (2.3)	8 (2.3)	7.9 (2.3)
Anxiety (change)	0.3 (0.6)	0.3 (0.7)	0.2 (0.8)	0.5 (0.7)
Anxiety (before)	1.8 (0.5)	1.9 (0.5)	1.9 (0.5)	1.7 (0.5)
Anxiety (during)	2.1 (0.5)	2.2 (0.6)	2.2 (0.7)	2.2 (0.6)
Process thinking	2.3 (0.8)	2.6 (0.9)	2.5 (0.8)	2.5 (0.9)
Time pressure	2.8 (1.0)	2.6 (1.0)	3 (0.8)	2.9 (0.9)
Explanation (word count)	39.3 (22.9)	82 (48.3)	33.1 (21.3)	73.0 (47.4)
Explanation (seconds)	133.9 (114.3)	289.9 (186.6)	112.6 (130.1)	247.8 (201.1)

Note: Columns 2 - 5 correspond to the four experimental conditions (control condition with no accountability, process accountability, outcome accountability, mixed accountability). The rows correspond to the various measures, namely basic demographics, two pre-matchstick arithmetic task measures, the manipulation check for the three accountability conditions, four measures based on the matchstick arithmetic task, five measures based on the post-task questionnaire, and two measures from the open-ended explanation/justification question concerning the subject's approach to the task.

Performance on the matchstick arithmetic task. The following analyses were designed to answer two main questions. First, is there evidence for meaningful differences in overall performance among the accountability conditions? Second, are any such differences moderated by problem type (A, B, C, and D, see Table 3.1)? To answer each question, we first examine a simple graphical display of the data and then assess each conclusion more rigorously by summarizing the results of Model 3.1.

[4] The postexperimental questionnaire also included an (open-ended) item that inquired about the expected (bonus) payment. It was included to check if the subjects doubted the (truthful) promise to receive a bonus. The great majority of subjects expected to receive a non-zero amount. Five subjects entered zero without further explanation and no one expressed any doubts about receiving the bonus.

To answer the first question, Figure 3.1 displays the total score (i.e. the number of correctly solved problems) for each subject in each condition. Visual inspection of the figure shows that performance was slightly worse in the process accountability condition relative to the other three conditions. One way to assess the magnitude of these differences is to estimate the probability that a randomly selected subject from one condition would outperform a randomly selected subject from another condition. We obtained these estimates using repeated random sampling and then used bootstrapping to compute the associated 95% uncertainty intervals. Zeroing in on the key process versus outcome accountability comparison, the probability that a randomly selected outcome-accountable subject would have a higher or equal total score than a randomly selected process-accountable subject equals 0.64 (95% uncertainty interval: [0.52, 0.76])[5]. Moving to our modeling analysis, the results of Model 3.1 are summarized in Table 3.3, which reports the estimated effects of the three accountability conditions relative to the control condition. Although the modeling results agree with the impression based on Figure 3.1, they also reveal considerable uncertainty in our estimates. This uncertainty makes strong conclusions unwarranted. For our key experimental comparison, the model assigns a probability of 0.91 to the hypothesis that process accountability lowers performance relative to outcome accountability.

Table 3.3: Basic results of Model 3.1 for Experiment 1

	Prior	Posterior		
		Mean	SD	P($\beta < 0$)
Intercept	Normal(0.5, 1)	0.48	0.41	
Process accountability	Normal(0, 0.5)	-0.30	0.22	0.92
Outcome accountability	Normal(0, 0.5)	0.04	0.23	0.44
Mixed accountability	Normal(0, 0.5)	-0.08	0.22	0.64

To answer the second question, Figure 3.2 reveals the percentage of subjects in each condition that correctly solved each problem. We observe that the pattern of results over different types of problems conforms quite well to the theoretical predictions of Knoblich et al. (1999). The difference in performance of process- and outcome-accountable subjects, however, does not seem obviously moderated by problem type. This conclusion is corroborated by the model-based analysis. For this analysis, we computed the difference in probability that a typical subject will solve a given problem if placed in a (a) process accountability condition versus (b) outcome accountability condition. These differences were then averaged according to problem type, resulting in a single estimate per problem type. We display the mean estimates and summarize the associated uncertainty in Figure 3.3. For each of the four problem types, the subject is expected to have a higher chance of correctly solving a given problem if placed in the outcome accountability condition. The four estimates are similar to each other (ranging from -7.3 to -3.8 percentage points) but highly uncertain, precluding any conclusion concerning the variation of the relative effect of process and outcome accountability over the different types of matchstick arithmetic problems.

[5]The probability that the randomly selected outcome-accountable subject will have a strictly higher (as opposed to higher or equal) total score is 0.54, 95% uncertainty interval: [0.41 – 0.66].

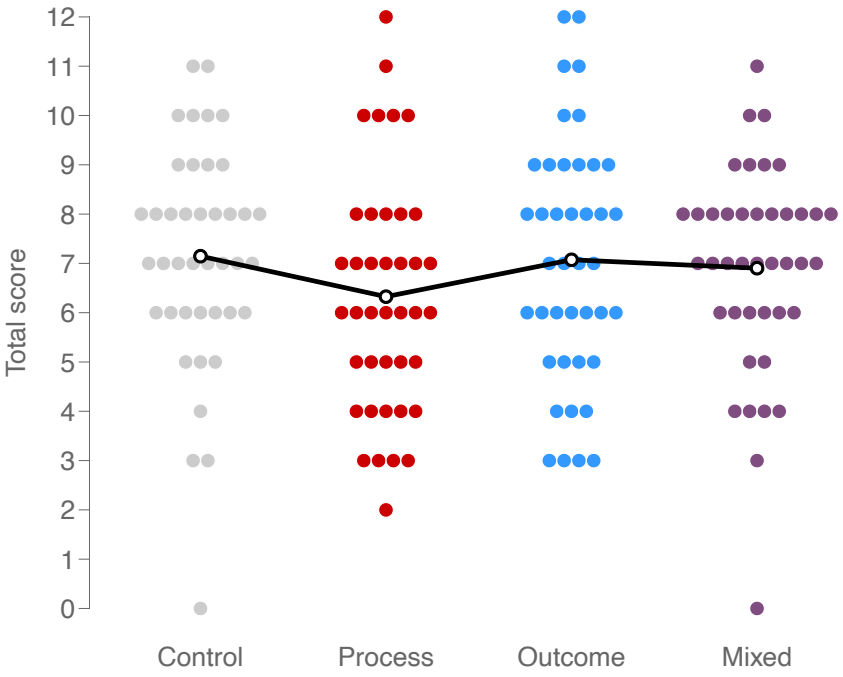


Figure 3.1: Performance on the matchstick arithmetic task in Experiment 1. The figure shows total scores (number of correctly solved problems, based on the strict scoring rule and using the full sample) across the four accountability conditions (control, process accountability, outcome accountability, mixed accountability). Each filled dot represents one subject. The points connected by the black line represent condition means. (Color version of this figure is available in the electronic version of this document.)

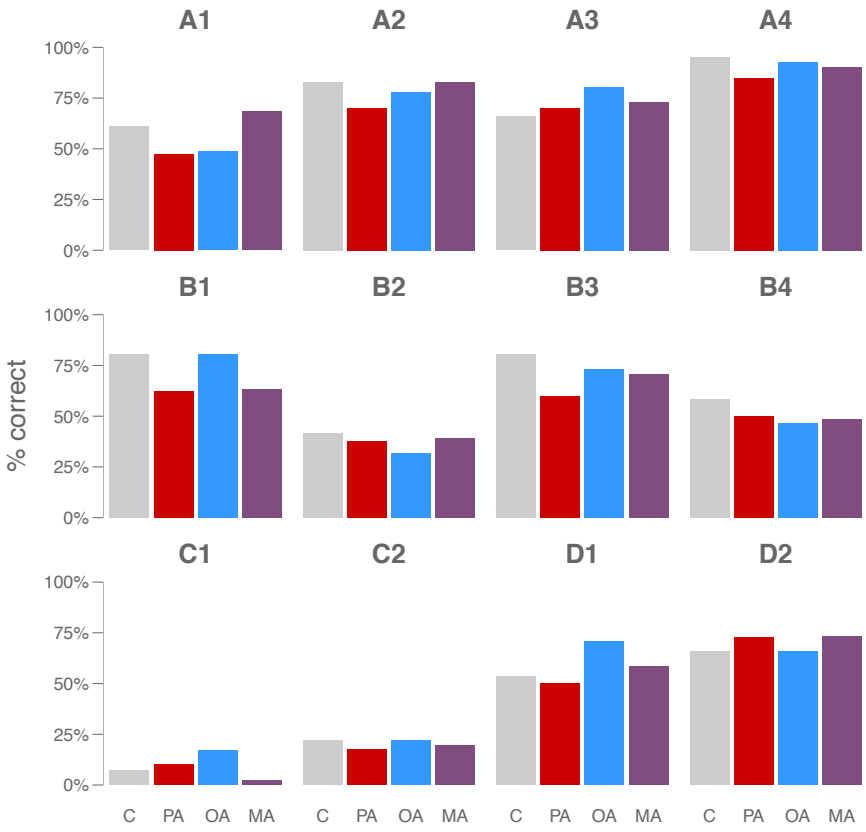


Figure 3.2: Performance on each problem of the matchstick arithmetic task in Experiment 1. The figure shows the percentage of subjects in each accountability condition (C = control, PA = process accountability, OA = outcome accountability, MA = mixed accountability) that correctly solved a given problem. (Color version of this figure is available in the electronic version of this document.)

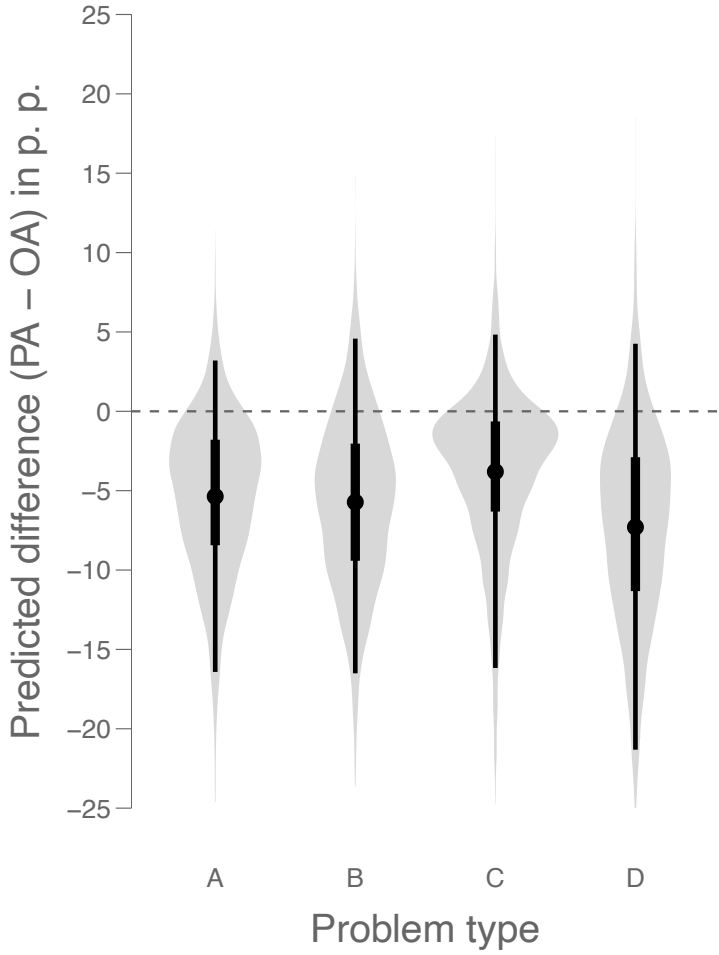


Figure 3.3: Predicted difference under process versus outcome accountability in the chance of solving a problem of a given type. The difference is expressed in percentage points, i.e. the chance under process accountability minus the chance under outcome accountability. The dots represent mean estimates, thick lines display 50% uncertainty intervals, thin lines display 95% uncertainty intervals, and the shaded areas summarize the entire posterior distributions.

Finally, we conducted additional analyses reported in the appendices. First, to assess the extent to which the results are dependent on specific modeling choices, we ran an extended analysis by constructing and fitting a series of eight models obtained by crossing two *scoring rules* (the original scoring rule versus a more lenient scoring rule, see *Scoring* in section *Methods*) with two *exclusion rules* (full sample versus a reduced sample that excludes accountability subjects failing the manipulation check, see *Preliminary analyses* in this section) and with two *estimation methods* (our main, Bayesian analysis versus a simplified classical analysis, see *Data analysis* in section 3.2). The results are reported in section *Appendix 2*. Overall, the variation in the results across these specifications is not major.

Second, a purely exploratory analysis revealed that although performance was on average lowest under process accountability among both male and female subjects, these effects were mostly driven by male subjects. We present and discuss the interactive effects of accountability and gender jointly for all our experiments in section *Appendix 5*.

Postexperimental questionnaire. Table 3.2 also reports the condition averages and standard deviations for all measures included in the postexperimental questionnaire. In general, we do not find any clear evidence for meaningful differences between the four conditions.

Discussion

The aim of Experiment 1 was to directly compare the effects of process and outcome accountability in an insight problem-solving task. To that end, we randomly assigned subjects to the control, process accountability, outcome accountability, or mixed acountability condition and had them solve a series of matchstick arithmetic problems (Knoblich et al., 1999). The results revealed that the average level of performance was lowest in the process accountability condition and similar in the other three conditions.

The finding that outcome-accountable subjects seemingly outperformed process-accountable ones stands in contrast to the dominant finding in this literature and could be taken as further corroboration of the hypothesis that the more deliberate thinking promoted by process accountability interferes with the processes that lead to insight (Eskenazi, 2015). But there are reasons to be skeptical of this conclusion. The first reason for skepticism is simply that the statistical evidence for the superiority of outcome accountability over process accountability is weak. To check the extent to which our results are dependent on the specific modeling choices we made, we explored a range of model specifications, in the spirit of a *multiverse analysis* (Steegen et al., 2016). Although the mean estimate for the effect of process accountability was consistently negative across these specifications, the associated uncertainty was also consistently large. Therefore, none of these analyses would suggest that the evidence for the superiority of outcome accountability over process accountability is strong. Furthermore, an exploratory analysis revealed that this effect was mostly driven by male subjects, raising the hypothesis that the relative effects of process and outcome accountability may be moderated by gender.

But even if were to accept that the performance advantage of outcome-accountable subjects over process-accountable ones in this study is not merely noise (in general or for male subjects

in particular), another exploratory analysis suggested an alternative interpretation of the result. In particular, we observed that, compared to the other three conditions, process-accountable subjects were more likely to terminate a trial early and subsequently *not* provide a candidate solution. A plausible interpretation of this pattern of behavior is that it indicates decreased solution-oriented motivation. Although the total number of such early terminations was low and therefore by itself cannot fully explain the difference in performance, it does suggest that the latter may not have anything to do with the assumed mechanism. Instead, the behavior of process-accountable subjects could reflect a reallocation of effort in response to the demand to justify one's approach as opposed to one's results. Put simply, process-accountable subjects may have tried less hard to find the solution.

Overall, although the current study provided some indication of the potential superiority of outcome over process accountability in the domain of insight problems (specifically, the matchstick arithmetic problems), the evidentiary and interpretative issues discussed above highlight a clear need for additional research.

3.3 Experiment 2

At first glance, Experiment 1 yielded evidence that process accountability results in worse performance on matchstick arithmetic problems than outcome accountability. This finding could indicate that the type of cognitive processing stimulated by process accountability is counterproductive for insight problems. But there are several issues with this conclusion and they motivate our next experiment.

First, the statistical evidence obtained in Experiment 1 concerning the performance difference between process and outcome accountability was weak. In addition, an exploratory analysis indicated that the finding may be moderated by the gender of the subject. We address these issues by replicating Experiment 1 while eliminating two experimental conditions in order to focus on the key comparison between process and outcome accountability.

Second, a closer look at the results of Experiment 1 suggested that process-accountable subjects may have simply had less motivation to accurately solve the matchstick arithmetic problems. We address this issue by modifying the design and by introducing additional measures.

Concerning the former (change in design), all subjects in Experiment 2 were given a monetary incentive to accurately solve the problems. This incentive provision can be thought of as a way to model the fact that, quite apart from accountability concerns, individuals typically have additional motivations to perform well. To ensure that the effect of the incentive scheme would not overshadow the effect of accountability, the incentive was weak. In particular, each correctly solved problem translated into a lottery ticket and the prize pool of the lottery included ten prizes of 10€ each.

Concerning the latter (additional measures), we used the two measures of thinking style introduced and validated by Novak and Hoffman (2009). These are the *rational* thinking style scale (relating to processing of Type 2) and the *experiential* thinking style scale (relating to

processing of Type 1). The former was used by de Langhe et al. (2011) to provide evidence of increased Type 2 processing under process accountability (relative to outcome accountability).

If Experiment 2 replicated a performance disadvantage of process-accountable subjects (of both genders) on the (most difficult problems of the) matchstick arithmetic task *and* if this result was mediated by higher Type 2 processing (as captured by the rational thinking style scale), the results would provide clear support for the theoretical account sketched out in the *Introduction*. Other experimental results (for example, no meaningful difference between the two accountability conditions) would increase the plausibility of alternative explanations of the results of Experiment 1.

Methods

Subjects. A total of 103 business students (enrolled in an undergraduate Management Accounting course) participated in the study. We excluded 11 of them because they either took part in a previous study using matchstick arithmeic problems ($n = 9$), failed the comprehension quiz (see *Procedure and materials*, $n = 1$), or failed to learn to interact with the study interface and consequently solved zero problems correctly ($n = 1$). The average age of the remaining 92 subjects was 20.8 years (SD = 1.1); 54 subjects were female and 38 subjects were male. A single session (which included an unrelated study that took part *after* this experiment) lasted about one hour and subjects were compensated for their time with course credit and a possibility to win an additional monetary prize (see *Procedure and materials*). Data collection took place in February 2020 (in total, 216 slots were made available).

Experimental design. Subjects were randomly assigned to either a *process accountability* ($n = 45$) or *outcome accountability* ($n = 47$) conditon.

Experimental manipulations. The accountability manipulations took place right before the main task. Subjects in the process accountability condition were informed that they were going to be asked to explain and justify their problem-solving approach after the problem-solving task. In contrast, subjects in the outcome accountability condition were informed that they were going to be asked to think about and evaluate their problem-solving performance. All subjects were informed that their answers (and, in the case of outcome accountability, their actual performance as well) were going to be displayed on the screen and reviewed by the researcher. They were also informed that the researcher was going to give them feedback and potentially ask additional questions.

Procedure and materials. The procedure was similar to that of Experiment 1, with a few changes introduced in order to improve the study and shorten its duration. After being welcomed and directed to individual cubicles, subjects reviewed the informed consent statement and began with the study. From this point onward, the procedure was fully computerized. Subjects first indicated their gender and birthyear and completed the 6-item measure of state anxiety (Marteau and Bekker, 1992). Next, they read the instructions, which also informed the subjects that a lottery would be conducted after the study and that 10 subjects would win a prize of 10€ each. The subjects were told, truthfully, that each correctly solved matchstick problem earns them one lottery ticket. The instructions concluded with a 10-item comprehension quiz. In

contrast to Experiment 1, each incorrectly answered question was returned to the end of the question queue and subjects needed to answer all questions correctly before proceeding. To prevent subjects from being stuck at this stage, the comprehension quiz ended if more than 10 errors were committed and the subject was later excluded from the study (this happened in one case, see *Subjects*). An additional change from Experiment 1 was that no training of Roman numerals took place. The training was deemed unneccessary because subjects were instead given a paper version of the instructions which, in addition to the main rules, listed Roman numerals from 1 to 15 and their Arabic counterparts. After the quiz, subjects completed the practice, read the manipulation text, and started the matchstick arithmetic task. We implemented two minor changes to the main task. First, the upper left corner featured a trial indicator (e.g. Round: 1 of 12). Second, after a trial was completed, subjects pressed a key to proceed to next trial (whereas in Experiment 1 the new trial began automatically). In every other way, the task was the same as in Experiment 1. After the main task, subjects again completed the measure of state anxiety (same as at the start, but this time the items were referring to how they felt while they were solving the problems) and the two scales (rational, experiential) of the thinking style questionnaire (Novak and Hoffman, 2009). A six-point response scale was used for these questions. Next, subjects were asked to describe their problem-solving approach and to predict how well they performed. Afterwards, the researcher reviewed the justification (performance) of process (outcome) accountable subjects, as specified in the manipulation text. Finally, subjects completed the second part of the session which was not related to this experiment.

Scoring. First, the submitted answers were automatically scored as correct or incorrect on the basis of *(i)* whether the correct matchstick was selected, *(ii)* whether its final position (x and y coordinates) was appropriate, and *(iii)* whether its orientation was appropriate. Only answers that fulfilled all three criteria were accepted as correct. To ensure that the scoring worked well, two additional steps were taken. First, the scoring rule was tested against old data which were scored by a human judge. Second, any solution that fulfilled all but one of the criteria was inspected. There were 37 such cases and in 21 cases, the answer was reclassified as correct on the basis of the inspection.

Data analysis. We used essentially the same approach as in *Experiment 1* for our main analysis. That is, we adapted Model 3.1 to reflect the fact that Experiment 2 only had two experimental conditions. The model used in Experiment 2 thus had a single dummy variable, with a value of 0 for outcome-accountable subjects and a value of 1 for process-accountable subjects. As in *Experiment 1*, we also attempted to fit the corresponding classical (maximum likelihood) model for our frequentist readers. Because the model did not converge, we fit a simpler model by dropping the varying slopes. Section *Appendix 3* reports these extended results.

Additional information. Additional information, including study materials and data, is available in the project's online repository (`https://osf.io/93gxw/`).

Results

Preliminary checks. We checked performance on the comprehension quiz and the number of early terminations. Statistics of both measures are available in Table 3.4. Similarly to

Experiment 1, 83.7% of the subjects made no more than one error on the comprehension quiz. For early terminations (defined as in Experiment 1), the average was smaller than in Experiment 1. Most importantly, the (positive) difference in the number of early terminations between the process accountability and outcome accountability conditions that we observed in Experiment 1 vanished in Experiment 2. In fact, the average was slightly (but not meaningfully) larger in the *outcome* accountability condition. For Experiment 2 we therefore do not find any evidence that subjects in the process accountability condition had lower solution-oriented motivation than subjects in the outcome accountability condition. Finally, we used the word count and response time of the *explain-your-approach* question as a manipulation check for process accountability. As in Experiment 1, process-accountable subjects on average expended more effort (at least as captured by these measures) on this question, suggesting that the manipulation was effective despite the fact that, unlike in Experiment 1, we did not tie the monetary bonus amount of process-accountable subjects to the quality of their explanations.

Table 3.4: Condition means (standard deviations) for various measures used in Experiment 2, separately for each experimental condition (see section *Appendix 1* for definitions).

	Process	Outcome
Gender (% female)	55.6	62.5
Age (years)	20.7 (0.8)	20.8 (1.3)
Comprehension quiz (errors)	0.7 (1)	0.9 (1.7)
Response time (seconds)	58.4 (12.1)	55.7 (17.7)
Early terminations	0.2 (0.6)	0.5 (1)
Total score	8.4 (1.9)	7.5 (2.8)
Anxiety (change)	0.6 (1)	0.8 (1.1)
Anxiety (before)	1.8 (0.7)	1.7 (0.7)
Anxiety (during)	2.4 (0.9)	2.5 (1.1)
Rational thinking	4.0 (0.9)	4.4 (0.8)
Experiential thinking	3.5 (0.8)	3.3 (0.8)
Rat. vs. exp. thinking	0.5 (1.3)	1.1 (1.1)
Explanation (word count)	85 (36.4)	65.2 (31.4)
Explanation (seconds)	244.9 (139.8)	159.2 (78.2)
Prediction (absolute)	7.8 (2.2)	6.9 (2.9)
Prediction (relative)	42 (18.3)	33.1 (18.7)
Optimism	-0.6 (1.1)	-0.7 (1.7)

Note: Columns 2 and 3 correspond to the two experimental conditions (process accountability, outcome accountability). The rows correspond to the various measures. A few clarifications: *(i)* the measures *Response time (seconds)*, *Early terminations*, and *Total score* are based on the matchstick arithmetic task (*Response time (seconds)* is based on the average number of seconds a subject used to think about a given matchstick arithmetic problem), *(ii)* three measures are based on other measures, namely *Anxiety (change)* (difference between the two anxiety scores), *Rat. vs. exp. thinking* (difference between rational and experiential thinking), and *Optimism* (difference between predicted and actual total score).

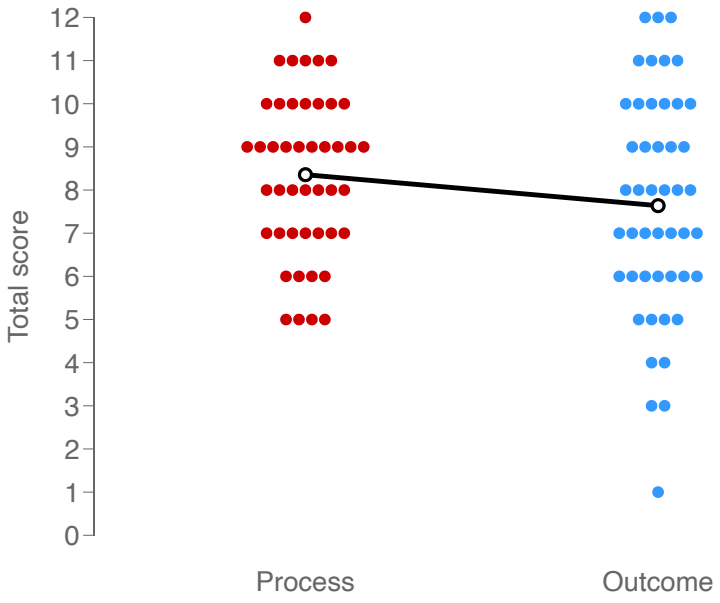


Figure 3.4: Performance on the matchstick arithmetic task in Experiment 2. The figure shows total scores across the two accountability conditions (process accountability, outcome accountability). Each filled dot represents one subject. The points connected by the black line represent condition means. (Color version of this figure is available in the electronic version of this document.)

Performance on the matchstick arithmetic task. To analyze the data from the matchstick arithmetic task we used essentially the same approach as in *Experiment 1*. Figure 3.4 shows the total scores across the two accountability conditions. The figure reveals that Experiment 2 does *not* replicate the poorer performance of process-accountable subjects observed in Experiment 1. In fact, the average performance is higher in the process accountability condition and the probability that a randomly selected *outcome* accountable subject would have a higher or equal total score compared to a randomly selected *process* accountable subject is less than 0.50 (0.48, bootstrapped 95% uncertainty interval: [0.36, 0.60])[6]. The associated modeling results are reported in Table 3.5. According to the model, the probability that process accountability reduces performance compared to outcome accountability is only 0.09. According to this analysis, the data are therefore consistent with a clear positive effect of process accountability as well as with an absence of a meaningful effect, but they are not consistent with a meaningful *negative* effect of process accountability.

Next, we checked if the difference in performance between the two accountability conditions

[6]The probability that the randomly selected outcome-accountable subject would have a strictly higher (as opposed to higher or equal) total score is 0.36, 95% uncertainty interval: [0.25 – 0.48].

Table 3.5: Basic results of (slightly modified) Model 3.1 for Experiment 2

	Prior	Posterior		
		Mean	SD	$P(\beta < 0)$
Intercept	Normal(0.5, 1)	0.84	0.40	
Process accountability	Normal(0, 0.5)	0.33	0.25	0.09

is moderated by problem type. Figure 3.5 displays the percentage of subjects in each condition correctly solving a given problem, and Figure 3.6 shows the model-based prediction for each problem type (see section *Results* in *Experiment 1* for an explanation of how these predictions were derived). On average, process-accountable subjects outperformed outcome-accountable subjects across the board, although the difference is essentially nil on the most difficult and most relevant problems (of type C). But, as in Experiment 1, the associated uncertainty is too large to admit any clear conclusion concerning moderation.

Finally, we checked for a possible interaction between accountability type and gender, following up on the results of an exploratory analysis in Experiment 1. Interestingly, the difference in performance between the two conditions in the current experiment was *again* driven by male subjects. But note that this means that the overall pattern of results was in fact just the opposite of that observed in Experiment 1: whereas male subjects in Experiment 1 performed better under outcome accountability than under process accountability, they performed better under process accountability than under outcome accountability in the current experiment. This and other observations make us skeptical of reading too much into these results. We present and discuss gender moderation effects jointly for all our experiments in section *Appendix 5* and also briefly address this issue in the *General discussion*. Importantly, other results of Experiment 2 are not meaningfully moderated by gender.

Thinking style. We computed two measures by averaging over the relevant items of the thinking style questionnaire: one corresponding to the *rational* thinking style scale and one corresponding to the *experiential* thinking style scale (Novak and Hoffman, 2009). We then sought to answer three main questions. First, do process-accountable subjects differ from outcome-accountable ones in their thinking style? Second, does thinking style correlate with performance on the matchstick arithmetic task? And finally, do differences in thinking style between the accountability conditions (assuming they exist) mediate the difference in performance?

Tentative answers to these questions are provided by Figure 3.7. First, in contrast to expectations and the results of de Langhe et al. (2011), the average score on the rational thinking style questionnaire is higher under *outcome* accountability. Indeed, the probability that a randomly selected outcome-accountable subject would have a higher or equal score on this measure compared to a randomly selected process-accountable subject is 0.67 (bootstrapped 95% uncertainty interval: [0.55, 0.78])[7]. (In both conditions, however, the scores for the rational thinking style are on average higher than the scores for the experiential thinking style.) Second, if anything, rational thinking style appears to *positively* predict performance on the matchstick

[7] The probability that the randomly selected outcome-accountable subject would have a strictly higher (as opposed to higher or equal) score on this measure is 0.63, 95% uncertainty interval: [0.51, 0.74].

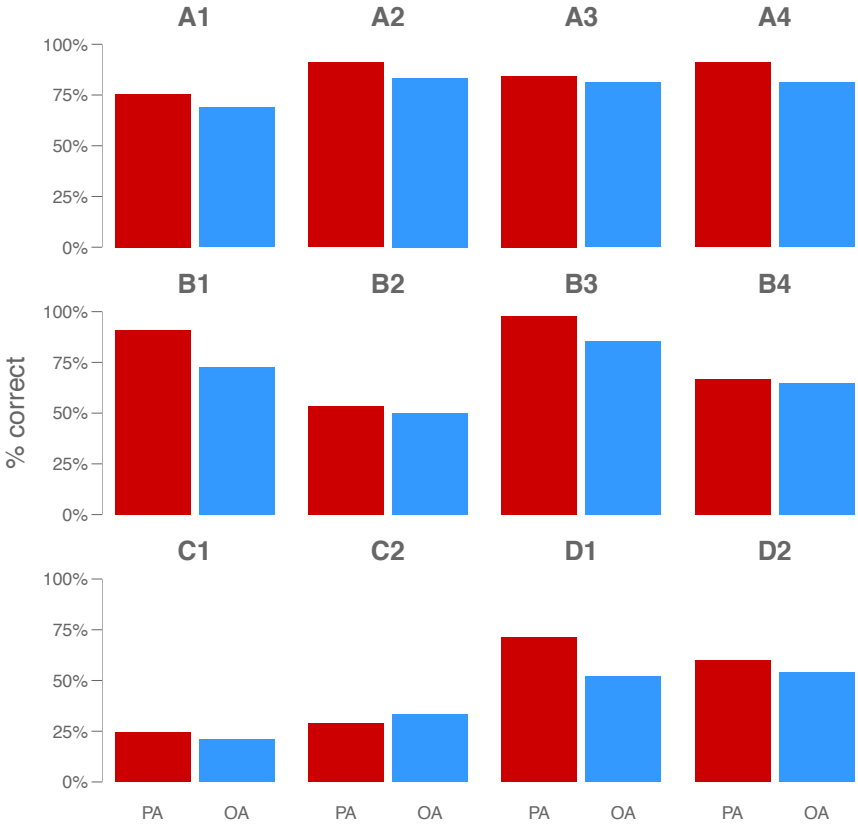


Figure 3.5: Performance on each problem of the matchstick arithmetic task in Experiment 2. The figure shows the percentage of subjects in each accountability condition (PA = process accountability, OA = outcome accountability) that correctly solved a given problem. (Color version of this figure is available in the electronic version of this document.)

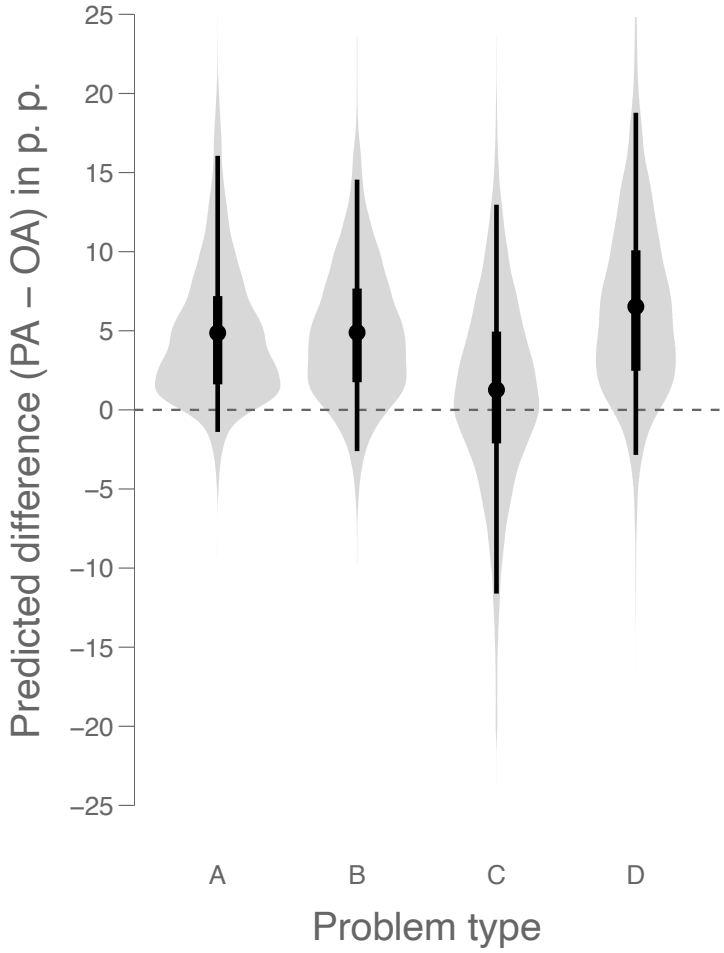


Figure 3.6: Predicted difference under process versus outcome accountability in the chance of solving a problem of a given type. The difference is expressed in percentage points, i.e. the chance under process accountability minus the chance under outcome accountability. The dots represent mean estimates, thick lines display 50% uncertainty intervals, thin lines display 95% uncertainty intervals, and the shaded areas summarize the entire posterior distributions.

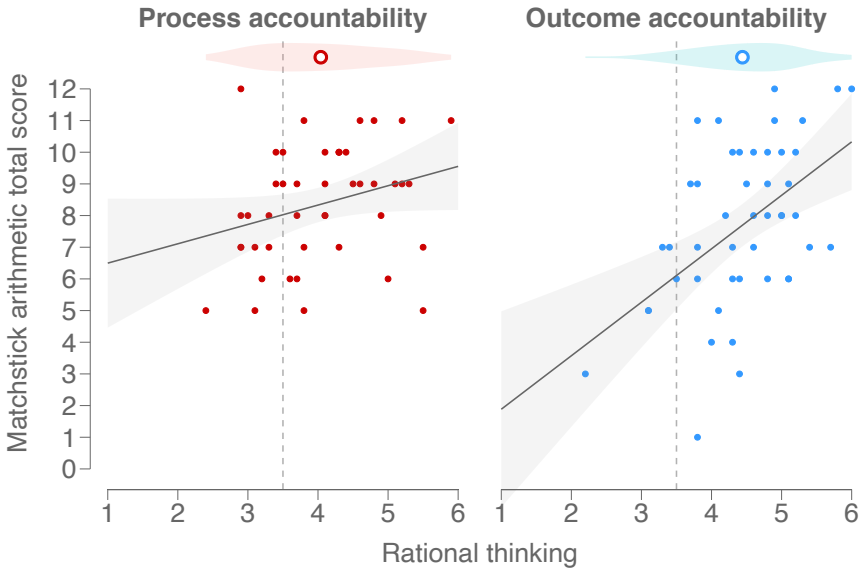
arithmetic task, although the correlation is not high (Pearson's $r = 0.35$, classical 95% uncertainty interval: [0.15, 0.52]; the correlation for the *experiential* thinking style is $r = -0.12$, 95% UI: [-0.32, 0.09]). These correlations are based on the *total* score of the matchstick arithmetic task, but a more fine-grained analysis does not show them to be driven solely by easier problems. For example, although it's true that the correlation between rational thinking style and performance is lower for problems of type D, the correlation for the most difficult problems (type C) is almost exactly the same as for the easiest problems (types A and B). And finally, the answer to our third question is that differences in thinking style clearly do not mediate the difference in performance on the matchstick arithmetic task between the accountability conditions, because outcome-accountable subjects score higher on rational thinking but perform worse on average. We return to this perplexing pattern of results in the *Discussion*.

Postexperimental questionnaire. Table 3.4 reports a few other measures based on the postexperimental questionnaire. In general, the results are similar for the process- and outcome accountability conditions. Most importantly, self-reported anxiety reliably increased during the task compared to the start of the session, and the average increase was similar for process and outcome-accountable subjects.

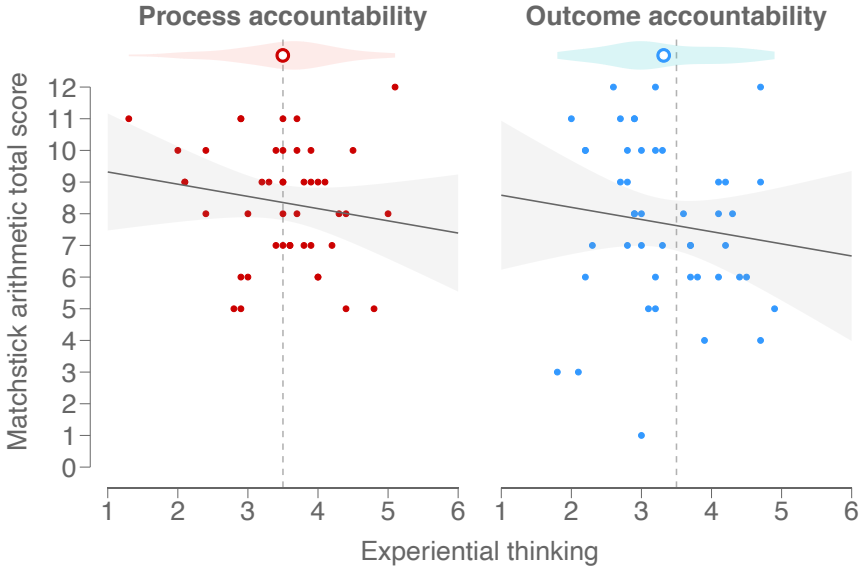
Discussion

The aim of Experiment 2 was to follow-up on the results of Experiment 1, in which we found some evidence that process accountability had a negative effect on performance on the matchstick arithmetic task compared to other experimental conditions, most importantly the outcome accountability condition. Apart from the uncertainty concerning the replicability (as well as possible gender-specificity) of this effect, the major motivation for the follow-up was the existence of two alternative accounts of this result. On one account, the lower performance of process-accountable subjects corroborates the hypothesis sketched out in the *Introduction*, which says that process accountability facilitates a shift towards more deliberate cognitive processing of Type 2 and that this shift interferes with processes that lead to insight (Wiley and Jarosz, 2012; Eskenazi, 2015). On the other account (supported by the results of an exploratory analysis), the key result of Experiment 1 was due to lower accuracy motivation on the part of process accounable subjects. To dissociate these two accounts, we modified the design of Experiment 2 in a way that ensured that subjects in all conditions had some motivation for accurately solving the problems, and also introduced additional measures of situation-specific thinking style (Novak and Hoffman, 2009).

Of these two accounts, the results of Experiment 2 clearly support the second one. The data we obtained in Experiment 2 are inconsistent with any meaningful negative effect of process accountability (relative to outcome accountability) on performance on the matchstick arithmetic task. In fact, performance was on average slightly *better* in the process accountability condition, although this result was driven by male subjects in particular. A further reason to be skeptical of the results of Experiment 1 is provided by the results of a supplementary experiment we conducted. In that experiment, we compared process accountability ($n = 114$) with a no accountability/control condition ($n = 110$) using the same matchstick arithmetic task as in Experiments 1 and 2. We found evidence for a *positive* effect of process accountability



(a) Rational thinking style and performance on the matchstick arithmetic task.



(b) Experiential thinking style and performance on the matchstick arithmetic task.

Figure 3.7: Relationship between thinking styles and performance on the matchstick arithmetic task. The circles (shaded areas) at the top display the average (density) of the variable on the *x*-axis, the grey best-fit lines are based on simple ordinary least squares models, and the light grey shaded areas represent the associated 95% uncertainty intervals. (Color version of this figure is available in the electronic version of this document.)

on performance ($\hat{\beta}$ = 0.34, 95% uncertainty interval: [0.01, 0.65], P(β < 0) = 0.02), which is consistent with the results of Experiment 2 but not with those of Experiment 1 (where process-accountable subjects performed worse than those in the no accountability/control condition).[8]

A clue as to why the prediction of poorer matchstick arithmetic task performance under process (relative to outcome) accountability did not pan out comes from the supplemental data provided by the two scales of the situation-specific thinking style questionnaire (Novak and Hoffman, 2009). These data suggest that a "rational" thinking style (measured using items such as "I tackled this task systematically" and "I was very focused on the steps involved in doing this task") *positively* predicts performance on the matchstick arithmetic task. This was true even for the most difficult problems — and these are of most interest because they are the ones most likely to require restructuring and insight. In contrast, the correlation for the "experiential" thinking style (measured using items such as "I relied on my sense of intuition" and "I had flashes of insight") had a weaker (and negative) correlation with performance. Together, these results contradict the assumption that (the most difficult) matchstick arithmetic problems are best tackled using a different, less deliberate approach. We return to this issue in the *General discussion*.

The analysis of the data from the rational thinking style questionnaire revealed another surprise: the average score was clearly higher in the *outcome* accountability condition. This finding conflicts directly with that of de Langhe et al. (2011) and indirectly with other studies (and current thinking) in the literature on process and outcome accountability. Together with the results mentioned above, this result also completes a curious pattern: rational thinking style positively predicted performance on the matchstick arithmetic task, but outcome-accountable subjects scored *higher* on this measure and *lower* on the matchstick arithmetic task. In fact, if we statistically adjust for rational thinking style (analysis not reported), the estimated effect of process accountability nearly doubles and the probability that it is positive increases to about 0.99. It is not entirely clear whether these results should be dismissed merely as noise or hint at some interesting underlying phenomenon. We return to this issue in the *General discussion* as well.

3.4 Experiment 3

In Experiment 3, we investigate the relative effects of process and outcome accountability on insight problem solving using a different task, namely the *remote associates* task (Mednick, 1962). In contrast to matchstick arithmetic problems, remote associate problems are verbal. Subjects are given a set of three initial words and instructed to find a fourth word that is associated with each of the three initial words. For example, often the fourth word forms a compound word or phrase when combined with a given initial word. This task tends to be difficult because of how the problems are constructed. The three initial words are always chosen so that the solution word is their *remote*, as opposed to *close*, associate (see Table 3.6).

[8] In this suppplementary experiment, performance was better under process accountability among both male and female subjects. If anything, the pattern was stronger among *female* subjects—see section *Appendix 5*.

Thus, the initial associations that subjects generate tend to be associated with only one or two of the three initial words and therefore do not qualify as a solution. Just like matchstick arithmetic problems, remote associate problems are among the most commonly used model problems in the literature on insight (Chu and MacGregor, 2011).

Eskenazi (2015) recently reported that subjects under process accountability perform worse on the remote associates task compared to subjects under a no accountability/control condition. In his studies, all subjects had a small monetary incentive to accurately solve the problems. Because of this design feature, the closest comparison for the studies of Eskenazi (2015) is our supplementary experiment, summarized in the *Discussion* section of *Experiment 2* (rather than Experiment 1). In this supplementary experiment, we found evidence that, if anything, process accountability seems to *improve* performance on the matchstick arithmetic task relative to a no accountability/control condition. Together, these results hint at a potentially interesting dissociation between the matchstick arithmetic task and the remote associates task. This dissociation is also of interest to the insight literature because both of these kinds of problems are frequently used in the investigations of insight.

The main aim of Experiment 3 is to provide the first direct comparison of the effects of process and outcome accountability on the ability to solve remote associates problems. This comparison provides another test of the hypothesis put forward in the *Introduction*. In addition, it helps us understand whether the presumed negative effect of process accountability on performance on this task is due to the specific demands of *process* accountability or a more general consequence of the social-evaluative threat that is also present in other forms of accountability. The second aim of Experiment 3 is to independently replicate the key finding of Eskenazi (2015, Chapter 3), namely worse performance under process accountability compared to *no* accountability. This seems like a crucial next step in this line of investigation, given that an important lesson of our earlier experiments is that accountability effects in this domain may be highly variable and sensitive to context. In addition, we earlier conducted a preliminary study in which we used a subtype of remote associate problems called *compound remote associates problems*. The results from this study yielded no evidence for a negative effect of process accountability (compared to a no accountability condition; outcome accountability was not included) and were in fact more consistent with a positive effect of process accountability. The study included only 22 subjects per condition, however, and differed in some other aspects from those of Eskenazi (2015). In contrast, Experiment 3 features over 100 subjects per condition and is essentially a close replication of Experiments 2 and 3 in Eskenazi (2015, Chapter 3).

Methods

Subjects. A total of 312 business students (172 female, 140 male) participated in the study. Their average age was 21.6 years (SD = 1.5). Participation in the study took about 30 minutes and subjects were compensated with course credit for their time. Data collection took place in January 2019 (in total, 350 slots were made available).

Experimental design. Subjects were randomly assigned to either a *no accountability* (n = 103), *process accountability* (n = 107), or *outcome accountability* (n = 102) condition.

Experimental manipulations. The accountability manipulations took place just before the start of the main task. All subjects were informed that people solve the [remote associates] task in different ways and that they should think about what their approach will be before starting. Subjects in the process accountability condition were additionally informed that they were going to be asked to explain and justify their approach after the task and that the researcher was going to review the explanation and might ask additional questions about it. In contrast, subjects in the outcome accountability condition were informed that their score (number of correct solutions) was going ot be computed and that the researcher was going to review their performance at the end and provide them with feedback about their level of performance.

Procedure and materials. The subjects were welcomed and directed to individual cubicles where they read the informed consent statement and began with the study. As in Experiments 1 and 2, the procedure was fully computerized. Subjects read the instructions for the remote associates task and completed three practice problems (with feedback). As in Experiment 2, they were provided with an incentive for correctly solving the problems. Specifically, each correct solution translated into a lottery ticket. The lottery was conducted after the data were collected and contained ten 20€ prizes. Following the manipulation text, subjects began with the study. We used 25 remote associate problems previously used by Eskenazi (2015) in Experiments 2 and 3 (Chapter 3). The problems are reproduced in Table 3.6. Subjects had 45 seconds to find the solution and 10 seconds to type it in. No feedback was provided at this stage. Following the task, subjects rated to what extent their approach to the task was intuitive or systematic. To save time, a single item was formulated for this purpose. Its text described the two modes of problem solving and asked the subject to use a visual analogue scale to place their own approach on the continuum between these two modes. Afterwards, subjects were asked to describe their problem-solving approach. Finally, they completed a short test of English vocabulary — LexTALE (Lemhöfer and Broersma, 2012) — before indicating their age and gender and answering several questions about their overall experience of the study (unrelated to the aims of this study). At the end, the experimenter reviewed the justification (performance) of process- (outcome-) accountable subjects, as specified in the manipulation text.

Scoring. Subjects' solutions were automatically scored as correct if they matched the prespecified solution noted in Table 3.6. In addition, all incorrect answers were manually reviewed. We found 111 (1.4%) answers which would have been correct if not for a spelling mistake or being in a plural form. Following Eskenazi (2015), we scored these answers as correct as well.

Data analysis. We used essentially the same approach as in *Experiment 1* for our main analysis, with three adjustments to the model. First, we adjusted the prior for the grand intercept to take into account prior knowledge about the higher difficulty of this set of remote associates problems compared to the matchstick arithmetic problems used in Experiments 1 and 2. The new prior is Normal(-1, 1) and corresponds to an expectation that for a typical subject facing a typical problem (and placed in the control condition), the probability of a correct answer would equal 0.27. Given that the subjects in the control conditions of Experiments 2 and 3 of Eskenazi (2015) (Chapter 3) correctly solved on average about 30% of problems, this seems appropriate. Moreover, the prior is only weakly informative, as the 95% uncertainty interval spans from 0.05 to 0.73 on the probability scale. Second, we included as predictors the standardized LexTALE score as well as its two interactions with the dummy indicators for process and outcome

Table 3.6: The remote associates problems used in Experiment 3

ID	Problem	Solution
1	broken, clear, eye	glass
2	coin, quick, spoon	silver
3	gold, stool, tender	bar
4	manners, round, tennis	table
5	playing, credit, report	card
6	rabbit, cloud, house	white
7	room, blood, salts	bath
8	salt, deep, foam	sea
9	chamber, staff, box	music
10	lick, sprinkle, mines	salt
11	pure, blue, fall	water
12	surprise, wrap, care	gift
13	barrel, root, belly	beer
14	strap, pocket, time	watch
15	walker, main, sweeper	street
16	chocolate, fortune, tin	cookie
17	big, leaf, shade	tree
18	board, magic, death	black
19	lapse, vivid, elephant	memory
21	rock, times, steel	hard
20	thread, pine, pain	needle
22	foot, collection, out	stamp
23	magic, plush, floor	carpet
24	shopping, washer, picture	window
25	stick, light, birthday	candle

From: Eskenazi (2015).

accountability. Finally, Experiment 3 did not feature a mixed accountability conditions, so the corresponding dummy variable was dropped. As in *Experiment 1* and *Experiment 2*, we also attempted to fit the corresponding classical (maximum likelihood) model for our frequentist readers. Again the model did not converge, so we fit a simpler model by dropping the varying slopes. Section *Appendix 4* reports these extended results.

Additional information. Additional information, including study materials and data, is available in the project's online repository (`https://osf.io/v8zwu/`).

Results

Preliminary checks. Table 3.7 reproduces the means and standard deviations of all measures, separately for each experimental condition. Three remarks are in order. First, we used the score on the LexTALE test as a measure of English proficiency. Because the majority of the subjects in this experiment were not native English speakers and because performance on the remote associates task critically depends on one's vocabulary, it is important to include such a measure. Indeed, we observe a clear positive correlation between the score on the LexTALE test and the number of correctly solved problems (Pearson's $r = 0.38$, classical 95% uncertainty interval: [0.28, 0.47]). Based on the gudelines presented by Lemhöfer and Broersma (2012), 56.7% of subjects were classified as B2 level (upper intermediate) users of English, with the rest classified as level B1 or lower. This classification is based on the LexTALE cut-off score of 60. Second, the number of early terminations (defined as in Experiments 1 and 2) is low and does not differ meaningfully between conditions. Therefore, similar to Experiment 2 but unlike Experiment 1, we do not find any evidence that process-accountable subjects had lower solution-oriented motivation. Finally, we compared the conditions with respect to the average effort expended in explaining one's approach to the task, as proxied by writing time and word count. Recall that the explanation was anticipated (and later reviewed by the researcher) for subjects in the process accountability condition but unanticipated for subjects in the other two conditions. As in Experiments 1 and 2, we observe that the means of both variables are highest under process accountability. In this experiment, however, the differences among conditions are substantially weaker and consistent with chance variation. What could explain this difference? Vis-à-vis Experiment 1, the quality of the explanation had no financial consequences for process-accountable subjects in this experiment. And vis-à-vis Experiment 2, process-accountable subjects in this experiment were not explicitly informed *right before* answering the question that they were about to answer the question that was going to be reviewed by the researcher. These methodological differences provide a plausible explanation of the results concerning the explanation word count and response time.

Performance on the remote associates task. Figure 3.8 displays the total scores of subjects in all three experimental conditions. Because of the importance of one's vocabulary for this task, we show the results separately for subjects whose English proficiency is high (level B2; left) and those whose English proficiency is low (level B1 or lower; right). Indeed, the results do seem to be moderated by English proficiency. Looking at the high-proficiency group in particular, the figure suggests that outcome accountability had a positive effect on performance compared to process accountability, with the subjects in the control condition somewhere in

Table 3.7: Condition means (standard deviations) for various measures used in Experiment 3, separately for each experimental condition (see section *Appendix 1* for definitions).

	Control	Process	Outcome
Gender (% female)	59.2	47.7	58.8
Age (years)	21.7 (1.8)	21.7 (1.3)	21.6 (1.5)
LexTALE	59.9 (8.6)	60.9 (7.7)	60 (8.4)
Total correct (RAT)	6.4 (3.3)	6.4 (3.1)	7 (3.5)
Response time (RAT)	24.8 (7)	24.9 (6.6)	26 (6.6)
Early terminations (RAT)	0.9 (2.6)	0.8 (2)	0.9 (2)
problem-solving approach	43.9 (23.4)	42.5 (23.4)	42.9 (23)
Explanation (word count)	53.3 (26.4)	58.3 (28.5)	53.1 (27.1)
Explanation (seconds)	128.7 (75.2)	134.4 (75.6)	125 (68.2)

Note: Columns 2 - 4 correspond to the three experimental conditions (control condition with no accountability, process accountability, outcome accountability). The rows correspond to the various measures, namely basic demographics, score on the LexTALE test, three measures based on the remote associates task, subjects' self-reported assessment of their problem-solving approach (where lower values correspond to a more intuitive approach and higher values correspond to a more deliberate, analytical approach), and two measures from the open-ended explanation question concerning the subject's approach to the task.

between. Still focusing on the high-proficiency group, we computed the probability that a randomly selected outcome-accountable subject would have a higher or equal total score than a randomly selected process-accountable subject to be 0.68 (bootstrapped 95% uncertainty interval: [0.58,0.78])[9]. The results are almost reversed if we examine the low-proficiency group. Among these subjects, a randomly selected *process* accountable subject would have a higher or equal total score than a randomly selected outcome-accountable subject with a probability 0.66 (95% UI: [0.54, 0.77])[10].

To investigate these findings more rigorously, we used a version of Model 3.1 (see *Data analysis* in section 3.3). The basic results are summarized in Table 3.8. Focusing on the comparison between process and outcome accountability, the model assigns a probability of 0.91 to the hypothesis that outcome accountability results in higher average performance than process accountability. This probability rises to 0.99 if we focus the prediction on a subject scoring half a standard deviation above the sample mean on the English proficiency test. Going in the other direction, the probability decreases to 0.61 for a subject scoring half a standard deviation below the sample mean, and to 0.30 for a subject scoring a full standard deviation below the mean. Overall, based on the assumption that the subjects with high English proficiency are the more relevant subgroup (as these are closer to native speakers), this analysis provides some support for the claim that outcome accountability is superior to process accountability when it

[9] The probability that a randomly selected outcome-accountable subject would have a strictly higher (as opposed to higher or equal) total score is 0.61, 95% uncertainty interval: [0.50, 0.71].

[10] The probability that a randomly selected process-accountable subject would have a strictly higher (as opposed to higher or equal) total score is 0.56, 95% uncertainty interval: [0.44 – 0.68].

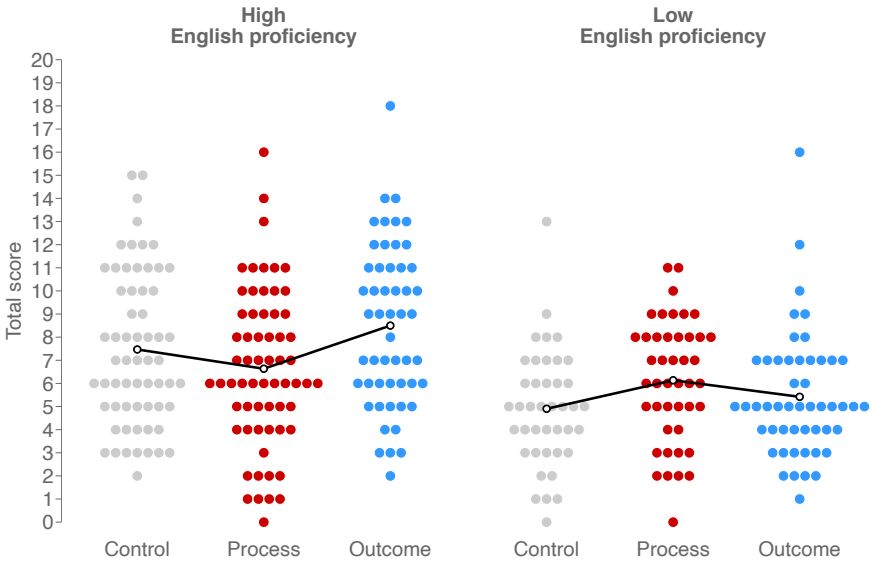


Figure 3.8: Performance on the remote associates task in Experiment 3. The figure shows total scores (number of correctly solved problems) across the three accountability conditions (control, process accountability, outcome accountability), separately for subjects with high (LexTALE score ≥ 60; left) and low (LexTALE score < 60; right) English proficiency. Each filled dot represents one subject. The points connected by the black line represent condition means. (Color version of this figure is available in the electronic version of this document.)

comes to performance on the remote associates task.

Table 3.8: Basic results of modified Model 3.1 for Experiment 3

	Prior	Posterior		
		Mean	SD	P($\beta < 0$)
Intercept	Normal(-1, 1)	-1.49	0.31	
LexTALE	Normal(0, 0.5)	0.37	0.08	0.00
Process accountability	Normal(0, 0.5)	-0.05	0.12	0.64
LexTALE × P. A.	Normal(0, 0.5)	-0.19	0.11	0.96
Outcome accountability	Normal(0, 0.5)	0.12	0.12	0.15
LexTALE × O. A.	Normal(0, 0.5)	0.07	0.11	0.27

Finally, as in Experiments 1 and 2, we checked for a possible moderating effect of gender. Fuller results are presented and discussed in section *Appendix 5*, but among the most relevant subgroup (subjects high in English proficiency), outcome-accountable subjects outperformed process-accountable ones among both genders.

Problem-solving approach. Figure 3.9 displays subjects' self-assessed problem-solving approach to the remote associates task, separately for subjects whose English proficiency is high (level B2; left) and those whose English proficiency is low (level B1 or lower; right). Lower values indicate a more passive, intuitive problem-solving approach whereas higher values indicate a more active and deliberate approach. Overall, the results show that the majority of subjects (58%) reported an approach that was more intuitive than deliberate (in contrast to the results for the matchstick arithmetic task). Comparing the scores across the three experimental conditions, the differences are minor. Thus, we do not find any evidence that process-accountable individuals would tend towards a more deliberate problem-solving approach compared to those in either the outcome accountability condition or the control condition. That is true even if we restrict the analysis to highly proficient English users. In fact, the change goes in the other direction. Among subjects with low English proficiency, a randomly selected *outcome*-accountable subject has the probability of 0.45 (bootstrapped 95% uncertainty interval: [0.33, 0.57]) of reporting a more or equally deliberate problem-solving approach than a random process-accountable subject. Among subjects with high English proficiency, the equivalent probability is 0.57 (95% UI: [0.46, 0.68]). Finally, we checked if problem-solving approach correlates with performance on the remote associates task. We find the correlation to be positive but low (Pearson's $r = 0.12$, classical 95% UI: [-0.01, 0.26]). This correlation is essentially the same for both English proficiency subgroups and there is also no interesting variation among the different remote associate problems. Thus, we do not find any evidence that a more deliberate approach is associated with worse performance on the remote associates task.

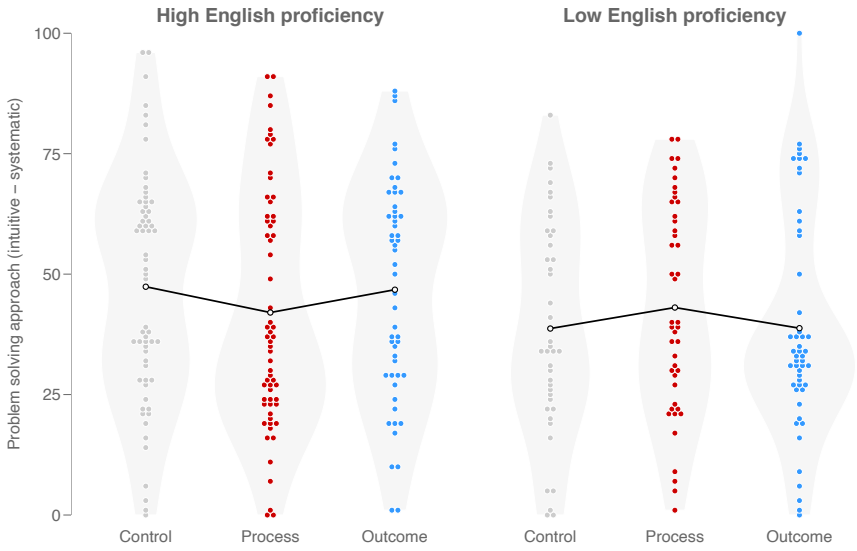


Figure 3.9: Self-reported problem-solving approach in Experiment 3. A score of 0 corresponds to a self-reported extremely intuitive approach to the remote associates task whereas a score of 100 corresponds to a self-reported extremely deliberate approach. Shown separately for subjects with high (LexTALE score ≥ 60; left) and low (LexTALE score < 60; right) English proficiency. Each filled dot represents one subject, the black line connects condition averages, and the grey outline tracks the density of the distribution. (Color version of this figure is available in the electronic version of this document.)

Discussion

In Experiment 3, we investigated the effects of process and outcome accountability on insight problem solving using the remote associates task. Prior research on insight problem solving suggested that process accountability would facilitate a more deliberate problem-solving approach that would prove detrimental to performance (Eskenazi, 2015). To investigate this idea, we randomly assigned subjects to a process accountability, outcome accountability, or no accountability (control) condition and had them complete 25 remote associates problems before asking them to self-assess their problem-solving approach.

We begin our summary of the results by focusing on the comparison between the process accountability and no accountability (control) condition. This comparison essentially provides a close replication of Experiments 2 and 3 in Eskenazi (2015, Chapter 3), as we have deliberately used the same problem set. In those experiments, Eskenazi (2015) observed a clear main effect of process accountability, with subjects in the control condition outperforming those in the process accountability condition. In contrast, the estimated effect of process accountability in the current experiment was essentially zero, providing further evidence of the variability of accountability effects in the domain of insight problem solving. At the same time, our

results do provide some evidence of an interaction, suggesting that process accountability did have a negative effect for more advanced English users. Because Eskenazi (2015) did not use LexTALE or a similar measure of English proficiency, we cannot make a direct comparison of the samples on this basis. It is possible, however, that his subjects were on average more proficient English speakers. One observation that is consistent with this conjecture is the fact that the average score of control subjects on the remote associates task was somewhat higher in the experiments of Eskenazi (2015) than in the current experiment (8.04 and 7.52 versus 6.40). Assuming that this conjecture is true, it brings our results closer together with those of Eskenazi (2015). It should be noted, however, that the preliminary experiment mentioned at the beginning of this section featured proficient English users (about 90% had a score of 60 or more on the LexTALE test) and yielded no evidence for a negative effect of process accountability. In fact, applying the same model as above resulted in the estimated posterior probability of about 0.90 to the hypothesis that the population effect of process accountability is *positive*.

Moving to the comparison between process and outcome accountability, we should first note that the results appeared to be moderated by English proficiency. In our view, however, the more relevant results are those of highly proficient English users. Among these subjects, we found evidence that outcome accountability facilitated better performance than process accountability. This result, if shown to be reliable, would provide an interesting contrast to the existing literature on process and outcome accountability, which has emphasized the superiority of process accountability.

What explains the better performance of outcome-accountable subjects compared to process-accountable ones? This question brings us to the second major dependent variable in the current study, namely the problem-solving approach used by the subjects. We used a simple self-report measure to test the idea that process accountability would facilitate a more deliberate, systematic approach that would prove maladaptive in the remote associates task. The results, however, do not provide any support for this idea. The correlation between a deliberate problem-solving approach was low and positive and process-accountable subjects did not report stronger reliance on a more active, deliberate problem-solving approach than subjects in the other two conditions. These results can of course be questioned on the grounds that the measure we have used was not properly validated, unlike the measures used in Experiment 2. Even if the validity of this measure is poor, however, the results are still interesting in light of the conjecture that process-accountable individuals not only rely on a more deliberate approach but also that they do so *because* it appears to be more justifiable (Eskenazi, 2015). To the extent that this is the case, we might still expect them to *report* a highly deliberate problem-solving approach, even if we question the idea that this report accurately tracks their actual problem-solving approach.

A clear limitation of the current experiment is that our sample consisted mostly of non-native speakers of English, although it should be noted that these are students who on average possess a solid command of English, as evidenced by their LexTALE scores and the fact that they are attending academic courses conducted in English. Still, extending our research by testing native speakers is a basic next step of this research program. Additional research is also needed because the statistical evidence for our conclusions is limited. In light of these observations, it would be inappropriate to make strong claims on the basis of the current experiment.

3.5 General discussion

Human progress depends on our ability to solve important problems in science, technology, business, governance, and art. Social science is in the position to contribute to this project directly but also on a meta-level, by investigating how we can facilitate successful problem solving among individuals and groups. Given that a lot of problem solving occurs in organizations and given the common need for accountability in organizational life, an important goal is to design accountability systems that promote the occurrence of fruitful insights that solve challenging problems. This overarching goal motivated the research reported in this chapter.

More specifically, we investigated the relative effects of process and outcome accountability in the domain of insight problem solving. Although previous academic research highlights the relative advantages of process accountability, the "special-process" theoretical view of insight and the results of Eskenazi (2015) suggest that in this domain, outcome accountability may in fact facilitate better performance than process accountability. The argument is that process accountability facilitates a problem-solving approach that is more reliant on the controlled, deliberate cognitive processing of Type 2 and that this approach interferes with implicit cognitive processing that would otherwise lead to restructuring and insight.

The results of three main and two supplementary laboratory experiments reveal this simple theoretical account to be inadequate. To structure this discussion, we first summarize the empirical results separately for the basic prediction (better performance under outcome accountability than under process accountability) and the two theoretical "building blocks" (the assumption that process accountability facilitates increased Type 2 processing relative to outcome accountability, and the assumption that such an increase in Type 2 processing interferes with restructuring and insight). Afterwards, we highlight the main takeaways of the research reported in this chapter and discuss their implications for future research on insight problem solving.

To test the basic prediction of better relative performance under outcome accountability, we used two well-known insight tasks. The first of these is the matchstick arithmetic task (Knoblich et al., 1999). In this quasi-mathematical task, automatically activated arithmetic and algebraic knowledge seems to place implicit constraints on the solution moves considered by the problem solver (Knoblich et al., 2001). For the more difficult problems (where these constraints prevent the discovery of the solution), this leads to an *impasse* and the cognitive representation needs to be restructured before the solution can be found. Our initial investigation (Experiment 1) suggested that the basic prediction of better performance under outcome (versus process) accountability might hold for these problems. But exploratory analyses and follow-up experiments suggested that this finding was unlikely to be robust and theoretically interesting. The data from two follow-up experiments were incompatible with a negative effect of process accountability on performance, either compared to outcome accountability (Experiment 2) or compared to no accountability (supplementary experiment, see the *Discussion* section of *Experiment 2*). We will note one potentially important design difference between the intial and the follow-up experiments when we consider the more general issue of the variability of accountability effects later in this discussion.

The second insight task we used is the remote associates task (Mednick, 1962). In this verbal task,

the initial solutions that individuals generate tend to be incorrect and potentially inhibit the retrieval of the correct solution. When individuals do manage to find the correct solution, they often report having obtained it with a sudden insight (Chu and MacGregor, 2011). Our results for this task (Experiment 3) were more in line with the basic prediction, with one major caveat. In particular, we found some evidence of an interaction between process accountability and subject's score on LexTALE, a test of English vocabulary (Lemhöfer and Broersma, 2012). Given that succesful performance on this task requires a well-developed English vocabulary, we deem the group of subjects who scored well on this test to be the more relevant one. Among these subjects only, performance was clearly better under outcome accountability than under process accountability. Although this result is suggestive, it needs to be independently replicated in a sample of native speakers. If replicated, it would point toward an aspect of creative problem solving that responds better to outcome accountability than to process accountability. As such, it would qualify the current academic view of process and outcome accountability and open up opportunities for future research.

Whereas the results concerning our basic prediction are mixed, the results concerning the broad theoretical assumptions that underpin this prediction are clear. In none of our experiments have we found any evidence for either of the two assumptions.

First, the correlations between task performance and measures designed to track the degree of deliberate, Type 2 cognitive processing were consistently low and *positive*. This was true for both of the tasks that we used. And it was true both for the validated measures used in Experiment 2 and the tailored measure employed in Experiment 3 (as well as in the supplementary experiment). These results are more in line with what has been termed the "business-as-usual" view of insight, according to which Type 2 processes that underlie successful performance on non-insight problems have an undiminished role in insight problem solving (MacGregor et al., 2001; Ball et al., 2015), than with the "special-process" view of insight.

Second, in none of the studies have process-accountable subjects reliably scored higher on measures designed to tap into the degree of Type 2 cognitive processing. The most surprising result was obtained in Experiment 2, where outcome-accountable subjects on average scored higher on the situation-specific rational thinking style questionnaire designed on the basis of dual-process theorizing (Novak and Hoffman, 2009). This finding is the exact opposite of the one reported by de Langhe et al. (2011), who found that process-accountable subjects scored higher on this measure. These findings are not necessarily incompatible. That is, both findings could be replicable and also *not* due to differences in accountability type manipulations, differences in the subject pool, etc. Even if we hypothetically could equalize all of these factors, the difference could still arise due to the different tasks used in the two studies. Clearly, the nature of cognitive processing employed depends on the task at hand. There is no *a priori* guarantee that a link between accountability type and cognitive processing will transplant from one task to another. But that does not mean that our result should be dismissed. The question remains: are these results replicable and, if so, what drives the difference? One *a priori* plausible candidate answer would be that process accountability is better at facilitating whichever type of cognitive processing is best suited for the task at hand. But as increased Type 2 thinking appeared to positively predict performance in Experiment 2, this hypothesis is not supported by the data.

What are we to make of all this? In our view, the first major takeaway is that there is a clear need for theory refinement. In particular, the crude dichotomy between Type 1 and Type 2 processes and the associated unidimensional conceptualization of their interplay appears insufficient in this context. It may therefore be fruitful to forego the coarse grouping of cognitive processes into Type 1 and Type 2 and focus on identifying and experimentally investigating the more specific cognitive processes (a) affected by accountability manipulations and (b) involved in insight. For example, Type 2 processing depends on working memory capacity but working memory itself can be broken down into various processes (for example, retrieval, response selection, inhibition of a predominant response, etc.) and components (for example, domain-general and domain-specific working memory). The same point applies to Type 1 processes — perhaps even more so, as this grouping encompasses an even wider variety of processes (Evans and Stanovich, 2013). The fruitfulness of this approach would likely by enhanced by also considering the *dynamic* or temporal component of insight problem solving. For example, insight problem solving may be conceptualized as involving the distinct stages of problem representation, active problem solving, *impasse* and restructuring. The most appropriate approaches, the associated cognitive processes, and the effects of different types of accountability may differ between these (or other) stages (DeCaro et al., 2016).

Another important takeaway from our studies concerns the overall variation in results that we observed. For example, the contrast between Experiment 1 and the two follow-up experiments featuring the matchstick arithmetic task may have stemmed from the use of a weak, probabilistic financial incentive for accuracy in the follow-up experiments. The relative effects of process and outcome accountability on insight problem solving may also depend on the task that is used, or the individual characteristics of the problem solver (such as their background abilities or even gender). We are not trying to suggest that any of these potential sources of variation is particularly important; the evidence for each is weak at best. The point is rather that the effects of accountability on insight problem solving appear to be highly heterogeneous. Indeed, the evidence for direct effects on performance was never clear-cut; the uncertainty intervals were always wide, even when the sample size was relatively large. For example, in the supplementary experiment there were more than 100 subjects in each condition and the posterior probability that the effect of process accountability was positive was about 0.98, but the lower end of the 95% uncertainty interval almost reached zero.

These observations lead to our final point that concerns accountability research in general. It is of course possible that the variation we observed is due to the particular domain we have investigated (insight problem solving), the nature of our accountability type manipulations, or both. But it is worth asking if the issue is perhaps not more general. After all, replications are rare in accountability research (see Chapter 2), reuse of measures is uncommon, and the materials used to manipulate accountability can vary quite a bit from study to study or are not carefully documented (see Chapter 1). Furthermore, by now it is well-known how flexible analyses, dichotomous thinking, selective publishing, and other characteristics of the current academic environment can lead to a highly biased, fragile knowledge base (Open Science Collaboration, 2015; Spellman, 2015). Add all of this together, and it becomes less clear which conclusions in the accountability literature are robust and which are not. Steps such as careful recording and sharing of experimental materials and data, attempts at standardization and reuse of measures and other study materials, and conducting independent replications are

therefore essential. All of these steps would facilitate the ultimate goal (the generation of insightful and genuinely predictive theory) by separating stable findings from the ones that are not and by exposing the gaps in the current theoretical understanding.

To conclude, the results of the studies reported in this chapter, alongside the earlier work of Eskenazi (2015) and Häusser et al. (2017), do not license any grand claims, but they do make a useful contribution to this research topic by highlighting some possibilities, challenges, and lessons that future research on accountability, insight, and creative problem solving ought to consider.

3.6 Appendix 1

Table 3.8: Variable definitions.

Label [Experiment]	Definition
Training [E1]	Number of correct responses in the training task administered before the main task (theoretical range 0 - 45).
Comprehension quiz [E1]	Number of correct responses on the quiz administered before the main task (theoretical range 0 - 6).
Manipulation check (% pass) [E1]	Correctly answering the manipulation check question, administered after the main task in the three accountability conditions.
Response time (seconds) [E1]	Average time spent thinking about the problem/searching for the solution.
Early terminations [E1]	Number of trials that were terminated at least 10 seconds before the time would run out *and* where no solution was then entered.
Total score [E1]	Number of correct answers using a "strict" scoring rule (only the solutions reported in Table 3.1 counted as correct).
Total score, alternative scoring [E1]	Number of correct answers using a more lenient scoring rule (with alternative solutions not explicitly forbidden by the instructions also counting as correct).
Anxiety (change) [E1]	Anxiety (during) - anxiety (before).
Anxiety (before) [E1]	Score (average, after recoding) on the state anxiety measure (Marteau and Bekker, 1992) administered before the main task.
Anxiety (during) [E1]	Score (average, after recoding) on the state anxiety measure (Marteau and Bekker, 1992) administered after the main task.
Process thinking [E1]	Average of three self-report items (administered after the main task) that asked about how much the subject thought about their approach to the task.
Time pressure [E1]	Score on a self-report item (administered after the main task) that asked about the experienced time pressure during the task.
Explanation (word count) [E1]	Number of words written in response to the *explain-your-approach* question.

Label [Experiment]	Definition
Explanation (seconds) [E1]	Time (in seconds) taken to submit the response to the *explain-your-approach* question.
Comprehension quiz (errors) [E2]	Number of errors on the quiz administered before the main task. Any incorrectly answered question was returned to the queue. Subjects who answered incorrectly more than ten times were excluded from the study, so the theoretical range on this measure for those that passed the quiz is 0 - 10.
Response time (seconds) [E2]	Average time spent thinking about the problem/searching for the solution.
Early terminations [E2]	Number of trials that were terminated at least 10 seconds before the time would run out *and* where no solution was then entered.
Total score [E2]	Number of correct answers using a "strict" scoring rule (only the solutions reported in Table 3.1 counted as correct).
Anxiety (change) [E2]	Anxiety (during) - anxiety (before).
Anxiety (before) [E2]	Score (average, after recoding) on the state anxiety measure (Marteau and Bekker, 1992) administered before the main task.
Anxiety (during) [E2]	Score (average, after recoding) on the state anxiety measure (Marteau and Bekker, 1992) administered after the main task.
Rational thinking [E2]	Score (average) on the rational situation-specific thinking style questionnaire of Novak and Hoffman (2009). A six-point response scale was used.
Experiential thinking [E2]	Score (average) on the experiential situation-specific thinking style questionnaire of Novak and Hoffman (2009). A six-point response scale was used.
Rat. vs. exp. thinking [E2]	The difference of the above two measures.
Explanation (word count) [E2]	Number of words written in response to the *explain-your-approach* question.
Explanation (seconds) [E2]	Time (in seconds) taken to submit the response to the *explain-your-approach* question.
Prediction (absolute) [E2]	Predicted number of correctly solved problems.

Label [Experiment]	Definition
Prediction (relative) [E2]	Predicted relative standing for the number of correctly solved problems (percentage of comparable students that woul score worse).
Optimism [E2]	The difference between Prediction (absolute) and Total score.
LexTALE [E3]	Score on the LexTALE English vocabulary test (theoretical range 0% - 100%). See Lemhöfer and Broersma (2012).
Total correct (RAT) [E3]	Number of correctly solved problems (theoretical range 0 - 25). Only the solutions reported in Table 3.6 counted as correct but small spelling mistakes or using the plural form did not disqualify an answer.
Response time (RAT) [E3]	Average time spent thinking about the problem/searching for the solution.
Early terminations (RAT) [E3]	Number of trials that were terminated at least 10 seconds before the time would run out *and* where no solution was then entered.
Problem-solving approach [E3]	Problem-solving approach self-assessed during the post-experimental questionnaire using a visual analogue scale, with 0 corresponding to an "extremely intuitive" approach and 100 corresponding to an "extremely systematic" approach. The two aproaches were described as is typically done in the literature on insight problem solving.
Explanation (word count) [E3]	Number of words written in response to the *explain-your-approach* question.
Explanation (seconds) [E3]	Time (in seconds) taken to submit the response to the *explain-your-approach* question.

3.7 Appendix 2

This appendix summarizes the results of an extended analysis of Experiment 1 data that was mentioned in the main text. The aim of this extended analysis was to check to what extent the results reported in the main text depend on the specific modeling choices we made. To that end, we constructed and fit a series of models defined by combination of (i) a *scoring rule* (the original scoring rule versus a more lenient scoring rule, see *Scoring* in section 3.2), (ii) an *exclusion rule* (full sample versus a reduced sample that excludes accountability subjects failing the manipulation check, see *Preliminary analyses* in this section), and (iii) an *estimation method* (our main, Bayesian analysis versus a simplified classical analysis, see *Data analysis* in section 3.2). Altogether, this yields eight models including the one that was reported in the main text.

The main results of all eight models are summarized in Figure 3.10. The figure shows the mean estimates and 95% uncertainty intervals for the three accountability parameters that capture the effect of each accountability condition relative to the control condition. The estimates printed in black correspond to Bayesian models and the estimated printed in grey correspond to models estimated using maximum likelihood estimation.

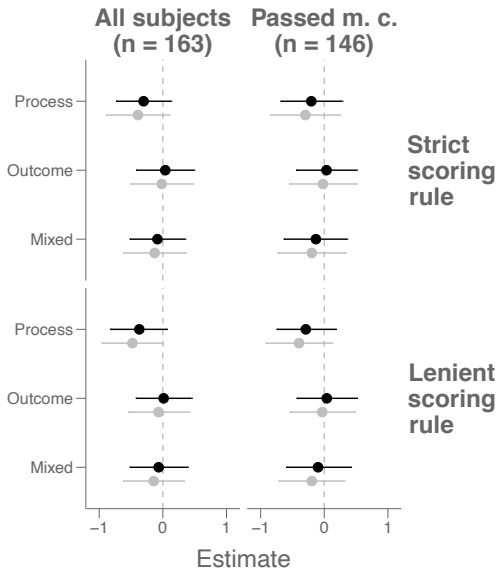


Figure 3.10: Multiple estimates (from eight separate models) of the effects of the three accountability conditions on performance on the matchstick arithmetic task in Experiment 1. The estimates printed in black are Bayesian ones, the estimates printed in grey are classical ones (maximum likelihood). Dots are estimates, lines are 95% uncertainty intervals. Passed m.c. = passed manipulation check.

3.8 Appendix 3

This appendix summarizes the results of an extended analysis of Experiment 2 data. We constructed and fit a series of models defined by combination of (i) an *exclusion rule* (reduced sample [see *Subjects* in section 3.3] versus full sample) and (ii) an *estimation method* (our main, Bayesian analysis versus a simplified classical analysis, see *Data analysis* in section 3.3). Altogether, this yields four models including the one that was reported in the main text.

Table 3.9 summarizes the results (the parameter corresponding to the process accountability dummy variable). It reports the point estimate (posterior mean for Bayesian models), the standard error (posterior standard deviation for Bayesian models), and the p value (posterior probability that the effect is negative for Bayesian models).

Table 3.9: Extended main analysis of Experiment 2.

Sample	Estimation method	Results		
		Estimate	SE	p
Main	Bayesian (informed priors)	0.33	0.25	0.09
Main	Maximum likelihood	0.38	0.25	0.14
Everyone	Bayesian (informed priors)	0.37	0.25	0.07
Everyone	Maximum likelihood	0.43	0.26	0.09

3.9 Appendix 4

This appendix contrasts the results of the main analysis of *Experiment 3* (using Bayesian estimation) with a roughly equivalent (see *Data analysis* in section 3.4) classical analysis (using maximum likelihood estimation).

Table 3.10 summarizes the results. It reports the point estimate (posterior mean for Bayesian models), the standard error (posterior standard deviation for Bayesian models), and the *p* value (posterior probability that the effect is negative for Bayesian models).

Table 3.10: Extended main analysis of Experiment 3.

Parameter	Estimation method	Results		
		Estimate	SE	*p*
Intercept	Bayesian (informed priors)	-1.49	0.31	
Intercept	Maximum likelihood	-1.54	0.31	
LexTALE (std.)	Bayesian (informed priors)	0.37	0.08	< 0.001
LexTALE (std.)	Maximum likelihood	0.38	0.08	< 0.001
Process Acc.	Bayesian (informed priors)	-0.05	0.12	0.64
Process Acc.	Maximum likelihood	-0.02	0.11	0.86
P.A. × LexTALE (std.)	Bayesian (informed priors)	-0.19	0.11	0.96
P.A. × LexTALE (std.)	Maximum likelihood	-0.21	0.11	0.06
Outcome Acc.	Bayesian (informed priors)	0.12	0.12	0.15
Outcome Acc.	Maximum likelihood	0.14	0.11	0.21
O.A. × LexTALE (std.)	Bayesian (informed priors)	0.07	0.11	0.27
O.A. × LexTALE (std.)	Maximum likelihood	0.05	0.11	0.62

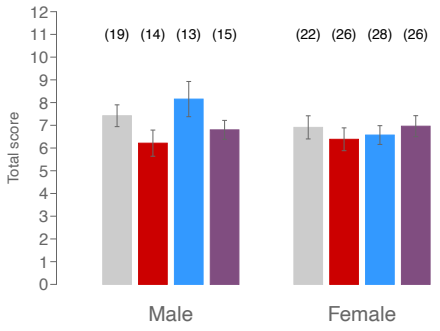
3.10 Appendix 5

This appendix considers the possible moderation of accountability effects by subject gender. Figure 3.11 shows the average score in each accountability condition, separately for each experiment and each gender.

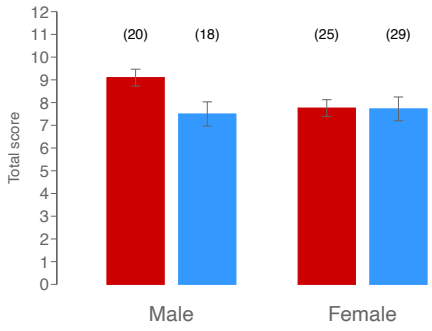
In Experiments 1 (Figure 3.11a) and 2 (Figure 3.11b), the difference between process and outcome accountability is mostly driven by the results of male subjects — but in different directions. In Experiment 1, male subjects are mostly responsible for the higher average score in the *outcome* accountability condition, whereas in Experiment 2 male subjects are responsible for the higher average score in the *process* accountability condition. Finally, in the supplementary experiment (Figure 3.11c; this experiment also featured the matchstick arithmetic task but only the *process accountability* and *no accountability* conditions), the average performance is higher under process accountability among both male and female subjects, and the difference is slightly larger among *female* subjects.

For Experiment 3 (which featured the remote associates task), we present the results separately for subjects high in English proficiency (LexTALE score ≥ 60; Figure 3.11e) and those low in English proficiency (LexTALE score < 60; Figure 3.11f). Among the more relevant subgroup of highly proficient English subjects and focusing on the key comparison, outcome-accountable subjects outperform process-accountable ones among both genders. Other results do indicate some moderating effects, for example the higher performance of process-accountable subjects that holds only for male subjects.

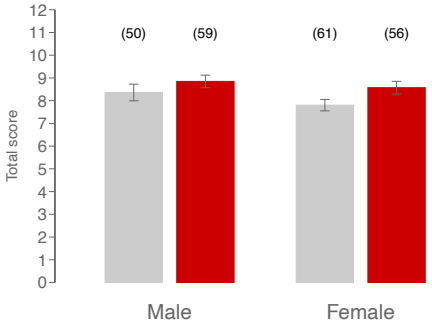
What can we conclude based on all this? Despite hints of moderation present in the data of almost all experiments, the results do not suggest a clear pattern overall. Of course, this does not mean that these effects don't "exist." But it does suggest they are highly variable, much like the main effects of accountability type that are of focal interest in this chapter. The conclusions and recommendations made in the discussion of variability in the *General discussion* therefore apply here as well.



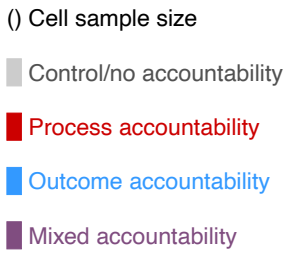
(a) Experiment 1



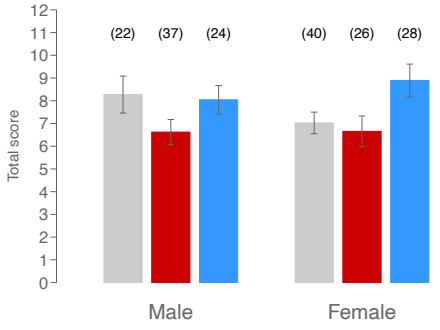
(b) Experiment 2



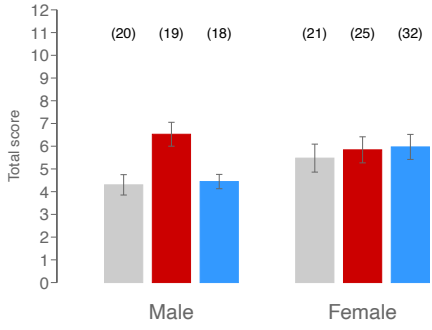
(c) Supplementary experiment



(d) Legend



(e) Experiment 3: High English proficiency



(f) Experiment 3: Low English proficiency

Figure 3.11: Accountability and gender. Bars correspond to averages, error lines to ± 1 standard error. (Color version of this figure is available in the electronic version of this document.)

Chapter 4

On adaptive decision making under process and outcome accountability

Do not choose a coward's explanation
That hides behind the cause and the effect
— Leonard Cohen favors outcome accountability

Summary:

Previous laboratory research on accountability suggests that holding individual decision makers accountable for their decision strategy or *process* typically results in better performance than holding them accountable for the eventual *outcomes* of their decisions. In the current chapter, we investigate if this holds true in dynamic decision environments in which decision makers face the need to combine reliable historical information and accuracy feedback. We hypothesize that the answer to this question depends on the kind of decision strategy that is best suited for a given decision environment. In particular, we suggest that the relative advantage of process accountability would reverse in a decision environment in which optimal decision accuracy is facilitated by the use of an information-frugal, heuristic strategy. We test this prediction in a laboratory 2 × 2 between-subjects experiment. In line with the prediction, we find that decision accuracy is higher under outcome accountability in such an environment. Surprisingly, however, we also find higher decision accuracy under outcome accountability in an environment in which optimal decision accuracy is facilitated by the use of an information-intensive strategy—that is, in a decision environment that, according to the current theory, should favor process accountability. This finding, along with the results of the analysis of supplementary self-report and process-tracing data, highlights the need for future research and suggests that, at least under certain conditions, outcome accountability supports more adaptive decision making.

Acknowledgments:

We thank Evelien Reusen for help in the recruitment of subjects and Stephan Kramer for commenting on an earlier version of this chapter.

4.1 Introduction

Organizations rely on a variety of formal and informal *accountability* mechanisms to influence the quality of decision making among their members. One family of approaches emphasizes the desired *outcomes* of decisions. For example, a common practice is to tie the individual's financial compensation to outcomes such as profit or customer satisfaction (Merchant and Otley, 2006). An alternative family of approaches emphasizes the quality of the *process* that was used for decision making. For example, decision makers could be asked to explain and justify their decision strategy as part of their performance review and evaluated on the basis of its soundness (Libby et al., 2004). Because both *outcome accountability* and *process accountability* can be used in the service of the same goal, an interesting and practically relevant question concerns their relative effectiveness in stimulating high quality decision making. To answer this question, a series of experimental studies in applied psychology, management accounting, auditing, personnel management, and related academic disciplines directly contrasted the two kinds of accountability and measured their effects on performance. The results of these studies indicate that individuals under process accountability typically outperform those under outcome accountability (Patil et al., 2014). However, the widespread adoption of outcome accountability systems in business practice raises the possibility that this literature overemphasizes the benefits of process accountability. In this chapter, we aim to help bridge academic research on accountability and the realities of business practice by investigating *adaptive decision making* under process and outcome accountability.

We use the term adaptive decision making to highlight the two key components of our experimental setting that often occur together in real-world decision making. The first key component is the presence of reliable historical information regarding the *cues* that are used to make the final choice. In the current study, the decision domain is choosing a location for oil-drilling from several candidate sites, so the cues are geological tests and the historical information concerns how accurate these tests have been in the past (*cue validity*). This historical information enables the decision maker to formulate and carry out a *decision strategy*: a way of using the cues to make a decision. For example, the decision maker could rely on historical information to weigh the importance of each cue, use these weights to determine the overall value of each alternative, and then choose the alternative with the highest overall value. In other decision domains, cues might be the results of personality tests (used in the HR domain to assess which job applicant would be best suited for a job), indicators of financial health (used in the commercial lending domain to assess the creditworthiness of a company), and so on.

The second key component is the presence of *feedback* from the environment. In our experimental setting, subjects receive clear outcome feedback informing them if the selected oil-drilling site turned out profitable or not. More generally, feedback can also occur in a partial or ambiguous form. The presence of outcome feedback significantly alters the decision-making situation. On the one hand, outcome feedback has the potential to facilitate the selection of decision strategies that are better suited to the decision environment. This is true for stable decision environments (Rieskamp and Otto, 2006) and becomes even more important in decision environments in which the underlying structure (and the appropriate decision strategy) changes in time (Bröder and Schiffer, 2006). On the other hand, the link between a decision

and its outcome is often characterized by a degree of randomness. In real-world decision making, even the best available decision strategies often have limited success (Gigerenzer, 2008). For this reason, outcome feedback could also hinder decision making if it led the decision maker to abandon an appropriate decision strategy.

Whereas we investigate the setting in which these two components co-occur, previous accountability research focused on investigating them separately[1]. Most studies focused on the first component, comparing the effects of process and outcome accountability in settings that lacked the feedback component. For example, Siegel-Jacobs and Yates (1996) investigated probability judgments in the legal domain, Brtek and Motowidlo (2002) studied judgments in the HR domain, and Dalla Via et al. (2018) explored the domain of project-funding decisions. In contrast, de Langhe et al. (2011) and Verwaeren (2019) focused on the second component but in a setting with no historical information. Although the findings from these two streams of literature overall support the same conclusion (superiority of process accountability over outcome accountability), we argue that they do not easily generalize to settings in which the two components co-occur, and that the relative superiority of process accountability would be diminished in these settings.

The aim of this research is therefore to investigate the relative ability of individuals under process or outcome accountability to use accuracy feedback to *adapt* their decision strategy to their *decision environment*. In this novel setting, we propose and test the hypothesis that this ability depends on the kind of decision strategy that is best suited for a given decision environment. In some environments, information-intensive strategies that attempt to integrate all available information yield the best performance. Surprisingly, however, in other environments simpler, heuristic strategies that ignore some of the information perform just as well or even better than the more complex, information-intensive strategies (Gigerenzer and Todd, 1999). For example, Luan et al. (2019) recently investigated the domain of hiring decisions using a real-world dataset from an airline company. A simulation analysis revealed that the heuristic strategy Δ-inference not only required less information but also made more accurate decisions than the information-intensive strategy logistic regression. The authors concluded that "under conditions of uncertainty common to managerial decision making, managers can actually make better decisions using *fast-and-frugal heuristics*."

We operationalize the two kinds of decision environments with reference to two prototypical and oft-studied decision strategies, the Weighted Additive rule (Payne et al., 1993) and the Take The Best heuristic (Gigerenzer and Goldstein, 1996). The first environment is designed in such a way that using the information-intensive Weighted Additive rule would yield high decision accuracy, whereas using the information-frugal Take The Best heuristic would yield substantially lower accuracy. For reasons that will become clear later (and to be consistent with prior research), we use the label *compensatory environment* for this first decision environment. In contrast, the second kind of environment is a mirror image of the compensatory environment: using the Take The Best heuristic would result in high decision accuracy, whereas using the Weighted Additive rule would result in substantially lower accuracy. Accordingly, we use the label *noncompensatory environment* for this second decision environment.

[1] An exception is Patil et al. (2017). But, as we explain later, this research considered a different form of process accountability than both our study and most other studies on the topic.

While a precise definition and detailed discussion of the Weighted Additive and Take The Best strategies will have to wait until the section *Decision environments*, it is important to note that the decision to base the two environments on these two strategies is not arbitrary but rather motivated by two different considerations. First, simulation analyses using real-world datasets reveal that these two strategies often perform very well and, depending on the environment, one strategy may be more suitable (or *ecologically rational*; Todd and Brighton, 2016) than the other (Czerlinski et al., 1999; Şimşek, 2013; Lee et al., 2017). Second, existing research suggests that these two strategies jointly explain much of human decision making and that individuals typically have a clear initial preference for one of them, although the direction of this preference depends on both task characteristics and subject population (Rieskamp and Otto, 2006; Heck et al., 2017; Garcia-Retamero and Dhami, 2009; Pachur and Marinello, 2013). When considering the ability of individuals to adapt to their decision environments, it is essential to take this initial preference into account. Because individuals tend to have a clear initial preference for either the Weighted Additive rule or the Take The Best heuristic and because we investigate environments in which one of these strategies is *ecologically rational* while the other one is not, we will observe how individuals under process and outcome accountability respond both when feedback mostly supports their initially preferred strategy and when it does not.

To summarize, in the current research we investigate adaptive decision making under process and outcome accountability in two kinds of decision environments. In the next two sections, we first summarize the empirical findings on process and outcome accountability and then define and discuss the two decision environments under investigation. Against this background, we formulate our exact research objectives and develop predictions in the final section of this *Introduction*.

Process and outcome accountability

Before summarizing the research on process and outcome accountability, let us clarify the scope of our review. We are specifically interested in studies that directly compared process accountability and outcome accountability and, more specifically, incorporated a *normatively ambiguous* form of process accountability (Patil et al., 2014). In this form, individuals are informed that they are to be evaluated on the quality of their *process*, such as their decision-making strategy or problem-solving approach. But they are *not* given specific guidelines (e.g. a standard operating procedure) on what constitutes an appropriate process. There are two reasons why we focus on this form of process accountability. First, it is the more generally applicable one, as it does not require specific knowledge about the appropriate strategy or process from the designer of the accountability system. Second, this is the form that has been typically found in the academic literature to yield better performance than outcome accountability[2].

[2] In contrast, when process accountability is paired with specific normative guidelines, process-accountable individuals tend to conform to this guidance and, as a result, their performance hinges on the appropriateness of the provided guidance for the problem at hand (Patil et al., 2017). In their study, Patil et al. (2017) informed their subjects that the company that employs them considers a specific decision rule to be "correct." When this decision rule in fact matched the structure of the environment, process-accountable subjects outperformed outcome-accountable ones. In contrast, when the rule was a poor fit for the environment, outcome-accountable subjects performed better.

We group the findings in this literature into three main clusters. Findings in the first cluster provide evidence that holding individuals process accountable elicits better performance than holding them outcome accountable. Relative to outcome accountability, process accountability has been reported to result in more accurate probability judgments in the legal domain (Siegel-Jacobs and Yates, 1996), more accurate judgments of leadership ability (Brtek and Motowidlo, 2002)[3], more accurate predictions in a linear multiple-cue learning task (de Langhe et al., 2011), better project funding decisions (Dalla Via et al., 2018), and higher profit in a business simulation task (Verwaeren, 2019). In addition to these straightforward performance measures, process accountability has been reported to be superior on other measures interpreted by the authors in normative terms. Relative to outcome accountability, process accountability has been found to decrease escalation of commitment to a failing course of action (Simonson and Staw, 1992)[4], self-serving behavior (Pitesa and Thau, 2013), and joint costs in a customer-supplier negotiation setting (Chang et al., 2013). Furthermore, process accountability has been found to enhance professional skepticism among auditors (Kim and Trotman, 2015) and exploration behavior in a business simulation task (Verwaeren, 2019).

Findings in the second cluster provide evidence on the underlying mechanisms that mediate the differences in performance of process- and outcome-accountable individuals. Methods used to provide this evidence include self-report (de Langhe et al., 2011; Häusser et al., 2017), covert observation of behavior by a confederate (Brtek and Motowidlo, 2002), process-tracing techniques such as information boards and eye-tracking (Siegel-Jacobs and Yates, 1996; Dalla Via et al., 2018), and analysis of supplementary data such as decision time and taped recordings (Häusser et al., 2017; Chang et al., 2013). Two main mechanistic hypotheses have been advanced. First, Siegel-Jacobs and Yates (1996) suggested that outcome accountability may be more likely than process accountability to result in stress, which in turn can have a negative effect on judgment or decision quality (Starcke and Brand, 2012). However, to date there is no direct corroborating evidence for this hypothesis. Häusser et al. (2017) did in fact find that accountability increases stress, but the increase was comparable under process and outcome accountability. Second, process accountability has been suggested to increase the *amount of information* considered in a given judgment or decision as well as the degree of *deliberate cognitive effort* applied to analyze the information (Siegel-Jacobs and Yates, 1996; Doney and Armstrong, 1996; Dalla Via et al., 2018). In line with this hypothesis, de Langhe et al. (2011) reports that process-accountable individuals scored higher on the situation-specific rational thinking questionnaire, which features items such as "I arrived at my answers by carefully assessing the information in front of me" and "I tackled this task systematically" (Novak and Hoffman, 2009). In a survey conducted among organizational buyers, Doney and Armstrong (1996) also found self-report evidence for a link between process (but not outcome) accountability and information search and analysis. Evidence not relying on self-report comes from Häusser et al. (2017), who found increased decision time under process accountability, and Dalla Via et al. (2018), who found a higher number of *fixations* (events, detected using an eye-tracking device, during which the eyes are relatively still, and which presumably reflect cognitive processing of the information that is attended to) under process accountability. Both

[3] This study appears to be more of a borderline case with respect to our "normative ambiguity" criterion. The subjects were apparerntly not informed about what the "correct" process would be, but the authors do mention the "structured format" that was given to the participants and which perhaps provided normative guidance.

[4] But see Schulz-Hardt et al. (2020) for a recently published failed attempt to replicate this effect

of these results may reflect increased search and/or more deliberate cognitive processing[5]. Finally, process-accountable subjects were rated as more attentive by a confederate in Brtek and Motowidlo (2002).

The final cluster of findings identifies boundary conditions for the superiority of process accountability over outcome accountability. In particular, de Langhe et al. (2011) found that the accuracy advantage of process-accountable subjects in a multiple-cue learning task vanished when the relationship between the cues and the criterion was non-linear (*making the task more difficult*), and Dalla Via et al. (2018) found similar decision quality in the two accountability conditions when the presentation format in a project funding decision featured a causal chain diagram (*making the task easier*). In a more creative domain, Häusser et al. (2017) found that ideas generated in the context of product development were similarly creative under process and outcome accountability. In addition to these boundary conditions related to the *task* or the *environment*, de Langhe et al. (2011) also found moderating evidence concerning the *stable individual characteristics* of the accountable individuals. In particular, the positive effect of process accountability originated from the subset of the subjects who were less able or inclined to think analytically. Overall, these findings make it clear that process accountability is not "consistently more desirable and uniformly superior to outcome accountability" (de Langhe et al., 2011), but they only show the *attenuation* of the typical finding rather than a complete reversal. The results of Davis et al. (2007) provide the most interesting exception; in a complex task designed to mimic military decision making, performance was higher under *outcome* accountability (although the average performance in this condition also deteriorated more with time than the performance of process-accountable subjects)[6]

Decision environments

As already noted, we operationalize two kinds of decision environments with reference to two specific decision strategies, the Weighted Additive rule and the Take The Best heuristic. We therefore begin this section by defining these two strategies. In the following, we use the standard terminology of research on probabilistic inference (e.g. Rieskamp, 2006). In particular, the goal of the decision maker is to select a decision *alternative* (in our experiment, a potential oil-drilling site) that is best according to some *criterion* (in our experiment, profit). The values of decision alternatives on this criterion are not directly known. But the decision maker can inspect one or more decision *cues* (in our experiment, results of geological tests) that correlate with the criterion. More specifically, we limit ourselves to the setting in which the cues are binary. The usefulness of these cues is operationalized as *cue validity*—the conditional probability that the cue will indicate the correct alternative, on the condition that the cue indicates a single alternative. In our experiment, these cues validities are explicitly communicated to the subjects as "historical information".

[5] A null result is reported by Siegel-Jacobs and Yates (1996), who found "no significant difference in the average number of cues used per judgment across conditions."

[6] There are two other exceptions: First, Slaughter et al. (2006) finds the *decoy effect* more pronounced under process accountability than under outcome accountability. But the statistical evidence is very weak. Moreover, the normative interpretation is not straightforward (see Arkes et al., 2016). Second, Eskenazi (2015) (Chapter 2) reports higher accuracy under outcome (versus process) accountability in an *affective* judgment task.

The Weighted Additive strategy is an information-intensive strategy that approximates normative models of decision making (Payne et al., 1993). When applied to a task of probabilistic inference, this strategy necessitates that all cue values be inspected, multiplied by their respective cue validity, and summed together to determine the overall value of a decision alternative. The alternative with the highest overall value is then selected (Bröder, 2000). Because cues lower in validity can potentially jointly overcome the influence of higher-validity cues, Weighted Additive is a *compensatory* strategy. In contrast, the information-frugal Take The Best strategy inspects the cues sequentially in order from the highest-validity cue to the lowest-validity cue. Crucially, the strategy stops the search immediately once a discriminating cue is found, and selects the decision alternative on the basis of that single cue (Gigerenzer and Goldstein, 1999). Because cues lower in validity can never overcome the influence of a high-validity cue, Take The Best is a *noncompensatory* strategy.

We construct the *compensatory environment* in such a way that the *Weighted Additive* is the more ecologically rational of the two strategies, and the *noncompensatory environment* in such a way that the *Take The Best* is the more ecologically rational of the two strategies. This approach has been fruitfully applied in a number of previous studies investigating human decision making. Whereas some researchers manipulate the ecological rationality of the two strategies by manipulating the costs of information acquisition (e.g. Bröder, 2003), we follow the approach under which the two strategies differ in their accuracy over the span of many decisions (e.g. Rieskamp and Otto, 2006). In particular, we constructed the compensatory environment in such a way that consistent application of the Weighted Additive rule would result in 80% of decisions being correct, whereas consistent application of the Take The Best heuristic would result in 60% of decisions being correct. In the noncompensatory environment, the opposite was the case—using the Take The Best heuristic would result in a decision accuracy of 80%, whereas using the Weighted Additive rule in a decision accuracy of 60%.

Several studies explored how people make decisions in these kinds of decision environments. For example, Bröder (2003) investigated the role of intelligence and working memory in adapting to these environments, Mata et al. (2010) compared the performance of younger and older adults, Kämmer et al. (2013) compared individual decision makers with dyads, and Rieskamp and Otto (2006) developed a formal theory of how individuals learn to adapt their decision strategy on the basis of accuracy feedback. Both theory and empirical findings highlight the role of individual's initial preference for one strategy over another. In laboratory studies in which the number of cues is low to moderate and there are no explicit costs for acquiring information, individuals typically have a clear initial preference for applying the Weighted Additive rule (e.g. Heck et al., 2017). In contrast, experienced decision makers seem to often prefer the Take The Best heuristic (Garcia-Retamero and Dhami, 2009; Pachur and Marinello, 2013). In our experimental setting, we assume that the Weighted Additive strategy would be more likely to be initially preferred.

Current research

The novel empirical work reported in this chapter consists of a 2 (accountability type) × 2 (environment type) between-subjects laboratory experiment. In addition to this experiment,

we also conducted a preliminary one (initial $n = 136$, n after exclusions = 127) which followed the same basic logic (a 2×2 between-subjects design with accountability type [process, outcome] and environment type [compensatory, noncompensatory] as factors). The results (and limitations) of this preliminary experiment are summarized throughout this section. The current experiment represents a confirmatory follow-up on this initial work. At the same time, it addresses some design limitations of the preliminary experiment and includes a considerably larger sample size.

Our work was designed to contribute to two related research objectives. In the following, we articulate these objectives and draw on existing literature on accountability and decision making to motivate our predictions.

Primary objective: *To compare decision accuracy under process and outcome accountability in compensatory and noncompensatory decision environments.*

Studies reviewed above suggest that performance is typically better under process accountability than under outcome accountability (Patil et al., 2014). Current theory and empirical evidence further suggest that this is because process accountability is more likely to facilitate an increase in the amount of information that is considered and the degree of deliberate cognitive effort that is applied to integrate the available information (Siegel-Jacobs and Yates, 1996; de Langhe et al., 2011; Dalla Via et al., 2018). These observations suggest that process-accountable individuals should be well-suited to making decisions in the compensatory decision environment, as defined in the preceding section. Older accountability literature also appears supportive of this conclusion (Lerner and Tetlock, 1999). For example subjects in Kahn and Baron (1995) indicated they would be more likely to use a compensatory decision strategy when making financial and medical choices if they needed to justify their decision, and Lee et al. (1999) found evidence for the same conclusion in a preferential choice setting (their manipulation also contained elements of outcome accountability, however). Because information-intensive and compensatory decision making results in high accuracy in the compensatory decision environment, these observations suggest that process-accountable individuals should perform well and likely outperform outcome-accountable individuals.

The more interesting question concerns the performance of process- and outcome-accountable individuals in the noncompensatory decision environment, in which accurate decision making is facilitated by the use of the information-frugal and noncompensatory Take The Best heuristic. On the one hand, the above discussion suggests that process-accountable individuals would be more likely to decide in an information-intensive, compensatory manner which would provide maladaptive in this environment. On the other hand, it is possible that the more deliberate approach of process-accountable individuals would help them grasp the underlying structure of the decision environment more quickly and allow them to flexibly adapt to the new conditions (De Dreu et al., 2006). Although existing theory is not strong enough to make a definite prediction, we argue that the former possibility is more likely. This prediction is further motivated by two additional observations. First, process accountability sometimes increases the *dilution effect* (Tetlock and Boettger, 1989), meaning that subjects' judgments were diluted more when irrelevant information was added to genuinely valid information (Nisbett et al., 1981). This result suggests that process-accountable individuals sometimes consider a larger amount of information even when it is less appropriate. In our noncompensatory

decision environment, accurate decision making is facilitated by the use of the Take The Best strategy, and this strategy sometimes requires that several cues are disregarded in favor of a single cue. Second, Siegel-Jacobs and Yates (1996) argue that process accountability increases *consistency* in the use of a judgmental (in our case, decision) strategy in the presence of outcome feedback. Although such consistency can be beneficial, it is likely to be detrimental in the noncompensatory environment, in which the optimal decision strategy (Take The Best) is unlikely to be the strategy process-accountable individuals initially adopt.

Altogether, the preceding discussion suggests that process accountability is likely to facilitate more accurate decision making than outcome accountability in the compensatory decision environment, whereas the opposite is likely to be true in the noncompensatory environment. To what extent do the results of our preliminary experiment conform to these predictions? We provide a graphical summary of results for all four experimental conditions in Figure 4.1. To show how decision accuracy changed throughout the task, we display a *moving average*[7]. In line with the prediction, we observe higher decision under *outcome* accountability in the noncompensatory environment (right). Surprisingly, however, decision accuracy is slightly higher under outcome accountability also in the compensatory environment (left), at least during the middle part of the task. In line with these observations, a formal statistical analysis provided evidence of a main (positive) effect of outcome accountability, qualified by an interaction with block (meaning that outcome accountability had a more positive effect later in the task). However, although the main effect was mostly driven by the noncompensatory environment, the evidence for the two-way interaction betweeen accountability type and environment type was too uncertain to conclude anything. Given this uncertainty, we deemed it important to conduct a second experiment. Another motivation for the second experiment stems from the way we implemented the task in the preliminary experiment. In particular, we used an information board on which the decision information (results of geological tests) was initially hidden *and* also implemented the constraint that only a single bit of information (one test result) could be inspected at a time. Although this design choice enabled us a more fine-grained view of predecisional behavior under process and outcome accountability, it can also be considered a limitation—clearly, the inability to observe all decision-relevant information at once changes the demands of the task at least to some extent (Lohse and Johnson, 1996). In the current experiment, we used a similar information board but removed the constraint that only a single bit of information could be inspected at a time. As a result, subjects could easily inspect all information at once. This solution allowed us to observe some aspects of predecisional behavior (see the secondary objective below) while at the same time presenting a more natural decision-making situation to the subjects.

Secondary objective: *To compare cognitive processing under process and outcome accountability.*

Although the relative effects of process and outcome accountability on decision accuracy are of primary interest, it is also important to consider the differences in cognitive or affective processing that give rise to differences in performance. Among other benefits, understanding the underlying mechanism(s) makes it easier to generalize predictions to novel contexts. In the current research, we rely on two additional data sources to shed light on cognitive processing under process and outcome accountability: *process-tracing* and *self-report*.

[7] The window of this moving average equals 52 trials, which is the length of one block.

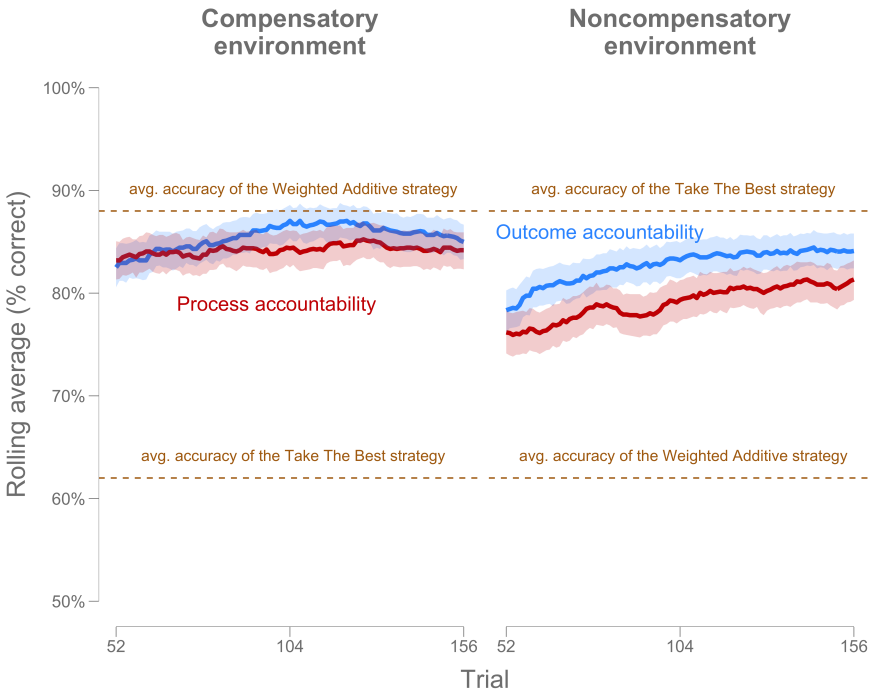


Figure 4.1: Accuracy (% of correct decisions) throughout the task in the preliminary experiment. The measure depicted on the *y*-axis is the moving average, computed using a window of 52 trials (length of one block). The lines denote condition averages, and the shaded areas display 95% uncertainty intervals obtained using a bootstrapping procedure (method: percentile-based, number of iterations: 1000). (Color version of this figure is available in the electronic version of this document.)

To trace subjects' predecisional behavior, the results of geological tests were initially hidden and had to be explicitly requested by clicking on the corresponding part of the screen. After being requested once, test results stayed visible for the entire duration of the trial. This setup allowed us to examine the *maximum* amount of information that was considered by the subject in a given decision, helping us test the prediction that process-accountable individuals would consider more information in their decision making (Siegel-Jacobs and Yates, 1996; Dalla Via et al., 2018), and also enabled some insight into how systematic and deliberate subjects' decision-making process was. As already noted, the preliminary experiment included a more intrusive version of this setup in which only one bit of information could be inspected at the time. The results showed that the total *unique* amount of requested information was similar under process and outcome accountability. But if the measure included all requests (including repeated ones), the average was reliably higher under process accountability. This result resonates with that of Dalla Via et al. (2018) and is potentially indicative of more deliberate decision making on the part of process-accountable individuals.

To complement the process-tracing results, we used self-report as our second additional source of data. In particular, we administered the situation-specific rational thinking style questionnaire (Novak and Hoffman, 2009) as part of the postexperimental questionnaire. Rational, deliberate thinking as captured by this instrument seems to match the reported beneficial effects of process accountability well, and indeed de Langhe et al. (2011) found that process-accountable subjects scored higher on this questionnaire. Our aim here was simply to replicate this finding. We have already used this measure in the preliminary experiment and found that it predicted accuracy more strongly in the compensatory than in the noncompensatory environment. Interestingly, we observed similar (and high) average score under process and outcome accountability. It seems plausible, however, that our setup, which allowed only a single bit of information to be inspected at a time, contributed to this pattern of results. Because we changed this aspect of study design, we employed the measure once again.

4.2 Methods

Subjects

The sample consists of 239 subjects recruited from a pool of business students enrolled in an undergraduate Management Accounting course [8]. The average age was 20.0 years (SD = 1.5), 139 subjects were female and 100 subjects were male. For their participation, subjects received course credit. In addition, some subjects received a financial bonus (see *Experimental manipulations*). Data collection took place in February 2020.

Experimental design

Subjects were randomly assigned to one of the four experimental conditions obtained by crossing the type of accountability (process, outcome) and the type of environment (compensatory, noncompensatory). There were 61 subjects in the condition process accountability × compensatory environment, 60 subjects in the condition outcome accountability × compensatory environment, and 59 subjects in each of the remaining two conditions.

Procedure and materials

Subjects (between one and eight per session) were jointly welcomed and escorted to individual cubicles. After reading the informed consent statement and agreeing to participate in the study, subjects indicated their age and gender and read the instructions. The latter introduced them to their role as a geologist working for an oil company, described the main task, and explained the key concept of cue validity (referred to as *test accuracy* in the study). The instructions also

[8] Overall, 241 subject participated in the study, but due to a technical error three subjects have incomplete data. For two of these subjects, the experimental script threw an error before the main task was completed, so they were excluded from the study. For the third subject, the error happened after the main task. This subject is included in the study but does not contribute any data to the analyses based on the data collected after the main task.

explained that each subject would be paired with another subject (recruited from Amazon MTurk), who would take the role of their manager. The subjects learned that the outcomes of their decisions would directly (financially) affect the "manager". Moreover, the "manager" would evaluate their performance and the subjects who earned the performance evaluation "exceeds expectations" would qualify for one of the 20 monetary bonuses of 25€ each (if the number of qualified subjects exceeded 20, random selection would be used). This setup provided the background for our accountability manipulations (see *Experimental manipulations*) and was carried out as described (i.e. there was no deception). The instructions also included two practice trials and were followed by the text of the accountability manipulation and the comprehension quiz. The latter included six questions that tested the subject's understanding of the instructions. Incorrect responses were followed by corrective feedback. After the quiz, subjects completed the main task, which consisted of four blocks of trials and is described below (see *Main task*). Afterwards, the subjects wrote an explanation of their decision-making approach and completed the postexperimental questionnaire. The latter consisted of 13 questions that all featured a six-point Likert scale. The first question asked "Do you believe you will be evaluated favorably by the manager?", the second question asked "Have you felt any responsibility for how your decisions would affect the second participant (the manager)?", and the third question asked "Would your decision making [*sic*] approach be different if decision outcomes directly affected you (as opposed to the manager)?". The variables based on these questions are referred to as *Optimism*, *Responsibility*, and *Different approach* in the section *Results*. The final ten questions were (minimally adapted) items from the rational subscale of the situation-specific thinking style questionnaire developed and validated by Novak and Hoffman (2009). The resulting variable (average of the ten answers) is referred to as *Rational thinking* in the section *Results*. Apart from the initial welcome, the procedure was fully computerized and programmed in Python using the PsychoPy package (Peirce, 2007).

Main task

The main task was a decision-making task consisting of 160 separate trials. In each trial, the subjects chose one of the three decision alternatives defined by a maximum of six cues. The subjects assumed the role of a geologist for this study (e.g. Rieskamp and Otto, 2006; Kämmer et al., 2013), so the three decision alternatives took the form of three potential oil-drilling sites (labeled A, B, and C) and the six cues took the form of six available geological tests. The subject's task was the same for every trial: to conduct one or more tests and then select the drilling site they predicted would result in the highest profit. Each decision would turn out either correct (resulting in a profit) or incorrect (resulting in a loss). As noted above, these outcomes (expressed in an experimental currrency) directly affected the second subject ("the manager") and not the subject who made the decisions.

The six geological tests included in the study were *chemical analysis*, *geophones*, *ground water analysis*, *gravimetry*, *microscopic analysis*, and *seismic analysis*. Two characteristics of each test were important in the study. First, the result of each test for a single oil-drilling site was either *positive* (indicating a higher chance that the site would be the most profitable one) or *negative* (indicating a lower chance that the site would be the most profitable one). Second, each test was associated with a percentage (labeled *Accuracy*) summarizing its track record.

Specifically, it indicated the percentage of past cases for which the test correctly predicted the most profitable oil-drilling site, conditional on the test results discriminating between the oil-drilling sites (i.e. cue validity). Subjects were informed that these percentages were based on a large amount of historical data and that they could therefore be trusted. (Indeed, the percentages were exactly right when considering all 160 trials of the task together.) The mapping between the tests and the percentages was determined randomly for each subject.

The study interface is similar to the one from Rieskamp and Otto (2006) and related work and is illustrated in Figure 4.2. Tthe initial positions of the tests were constant for each subjects but varied randomly among subjects. At the start of a trial, no test results were displayed. Subject could conduct each test (free of cost) by clicking on the corresponding icon. After the click, the test results for all three oil-drilling sites appeared on the screen and stayed there for the duration of the trial. Thus, subject could conduct between zero and six tests in each trial before making a decision. In addition to the decision, the experimental script recorded which tests were conducted and when.

There were four blocks in total, separated by short breaks. Each block consisted of 40 trials. In a deeper sense, the set of 40 trials was the same in each block. But the order of the trials and the mapping between the three underlying decision alternatives and the three oil-drilling sites was randomly determined for each subject and each block. The exact nature of the trials differed between the two decision environments and is explained in the section *Experimental manipulations*. In the first block, a decision was immediately followed by the next trial. In contrast, in blocks 2 - 4 a decision was followed by immediate feedback, i.e. the subjects were notified whether their decision turned out correct or not. Subjects were informed in the instructions that they could use this information to improve their accuracy.

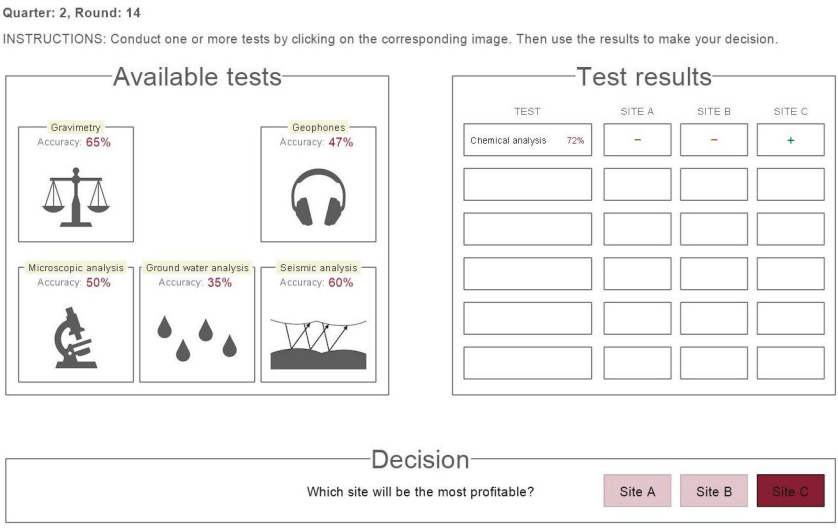


Figure 4.2: Screenshot of one trial in the main task. (Color version of this figure is available in the electronic version of this document.)

Experimental manipulations

Accountability type. The operationalization of process and outcome accountability built upon the setup outlined in the section *Procedure and materials*. Subjects were informed that the outcomes of their decisions would directly (financially) affect the second subject (who would be assigned to the role of their manager), that this second subject would evaluate their performance, and that their chance of earning a financial bonus of 25€ would be directly tied to the evaluation they received. More specifically, the second subject would evaluate their performance as either "below expectations", "meets expectations", or "exceeds expectations," and among the subjects who received the evaluation "exceeds expectations," 20 would be randomly selected to receive the bonus of 25€ each. Subjects in all conditions were also informed that the "manager" would follow a "standardized procedure" when evaluating their performance and that this procedure would involve reading the task instructions and reviewing the performance of several other subjects in the study. The difference between the manipulations of process and outcome accountability was in what the "focus" of the "manager" would be and in the information they would be able to access. Subjects in the *process* accountability condition were informed that the "manager" would focus on their decision-making strategy and that "strategy" refers to the method used to make decisions on the basis of available tests. They were also informed they would have to "explain and justify" their decision-making strategy after the main task, and that the "manager" would review (i) their explanation/justification and (ii) several decisions randomly selected from each "quarter" (block). When reviewing the latter, the "manager" would be able to see which tests were conducted, what their results were, and what the subject's decision was — but not the outcome of the decision. In contrast, subjects in the *outcome* accountability condition were informed that the "manager" would focus on the outcomes of their decisions and that "outcome" refers to whether a decision turned out a success (i.e. correct) or not. They were also informed that several decisions in each quarter would be randomly selected and that their outcomes would be reviewed by the "manager." The latter would only be able to see what the outcome of a given decision was and not "how" the subject made the decision. That is, the manager would not be able to see which tests were conducted, what their results were, and which drilling site was selected by the subject.

Decision environment. To operationalize the two types of decision environment, we manipulated the trials that the subjects encountered during the main task. We prepared two sets of 40 unique trials, one for the *compensatory* decision environment and one for the *noncompensatory* decision environment. These two sets of trials (from now on: environments) did not differ in their surface-level representation (oil-drilling sites, geological tests). Instead, the key difference was in the long-run performance of the two prototypical decision strategies. In the compensatory environment, consistent application of the (compensatory) Weighted Additive strategy would result in 80% of correct decisions, whereas consistent application of the (noncompensatory) Take The Best strategy would result in 60% of correct decisions. And the noncompensatory environment was the mirror image of the compensatory one: the Take The Best strategy would yield 80% of correct decisions, and the Weighted Additive strategy would yield 60% of correct decisions[9]. In terms used in the *Introduction*, the Weighted Additive

[9] The accuracy of the Equal Weights strategy was 65% in the compensatory environment and 47.5% in the noncompensatory environment. The Equal Weights strategy is a special case of the Weighted Additive strategy with all weights set to the same constant.

strategy was ecologically rational in the compensatory environment, and the Take The Best strategy was ecologically rational in the noncompensatory environment. Our aim in constructing the two environments was to minimize all differences between the environments save for this crucial property. To accomplish this task, we represented the combined environments as a four-dimensional 40 × 6 × 3 × 2 arrray of ones and zeroes (trials × cues × alternatives × environments) and programmed a custom genetic algorithm to find an appropriate solution. The algorithm's fitness function facilitated the evolution of several desirable properties. For example, we required that the trials and cues be unique, that the cues be diagnostic (cue validity > 0.33) but not perfectly so (cue validity < 1.00), that the cue validities communicated to the subjects would in fact be exactly true for a given environment, that the two key strategies would always make unambigous predictions and that the overlap between their predictions was 50% in each of the environments, and so on. The final solution is fully reproduced in section *Appendix 1*. The cue validities of the six cues were [0.65, 0.53, 0.50, 0.46, 0.44, 0.35] in the compensatory environment, and [0.72, 0.65, 0.60, 0.50, 0.47, 0.35] in the noncompensatory environment.

Data analysis

We used Bayesian estimation for modeling purposes. The models were fit using the R package `brms` (v2.4.0, Bürkner, 2017), which serves as an interface to the `Stan` language (Gelman et al., 2015). Our general approach was to obtain four values for each subject (one for each block) by averaging over the relevant trials. We then regressed these values on the fixed effects of environment type, accountability type, block, and all their two-way and three-way interactions — while also including a varying intercept (random effect) for subjects.

To obtain the estimates that could be interpreted as ANOVA-style main effects and interactions, we used sum-to-zero coding for all predictors. For decision environment, we coded noncompensatory environment as $+\frac{1}{2}$ and compensatory environment as $-\frac{1}{2}$. This is reported as Environment (non. vs. comp.) in the table summarizing the modeling results. Similarly, we coded process accountability as $+\frac{1}{2}$ and outcome accountability as $-\frac{1}{2}$ (Accountability (pro. vs. out.)). Finally, there were four blocks, so we created three block predictors. These are Block-A (2/3/4 vs. 1) (Block 1: $-\frac{3}{4}$, Blocks 2/3/4: $+\frac{1}{4}$), Block-B (3 vs. 2) (Block 3: $+\frac{1}{2}$, Block 2: $-\frac{1}{2}$, Blocks 1/4: 0), and Block-C (4 vs. 2/3) (Block 4: $+\frac{2}{3}$, Blocks 2/3: $-\frac{1}{3}$, Block 1: 0). Thus the intercept represents the grand mean and the parameter estimates represent the difference identified in the label of the predictor.

Depending on the nature of the dependent variable, we used different response distributions and link functions. If the latter are not explicitly identified for a given analysis, it means that the normal distribution with an identity link function was used. For convenience, we used the default (weakly informative) priors supplied by the `brms` package.

When presenting the results in tabular form, we report posterior means and standard deviations. In addition, we use asterisks to highlight those parameters for which the posterior probability that the effect is in a particular direction (i.e. positive or negative) is greater than or equal to 0.95 (one asterisk), 0.975 (two asterisks) or 0.995 (three asterisks). These approximately correspond to classical two-tailed significance levels of 0.10, 0.05, and 0.01.

When presenting the results in the text, we report 95% uncertainty intervals (percentile-based) and/or the posterior probability that the estimate is of a given sign.

For the benefit of frequentist readers, we fit the equivalent version of our main models using classical maximum likelihood estimation. These results are reported in section *Appendix 3* (alongside other supplementary analyses).

Additional information

Additional information, including study materials and data, is available in the project's online repository (https://osf.io/zrhm7/).

4.3 Results

Preliminary analyses

Basic descriptive statistics are reported in Table 4.1 and Figure 4.3. The table reports, separately for each experimental condition, the means and standard deviations of the measures used in the study, and the figure displays the correlations among the variables. The latter include the measures as well as the two manipulated variables (type of accountability and type of environment).

To check subjects' understanding of the instructions, we examined the number of errors made on the comprehension quiz that was administered right before the main task. Overall, the majority (80.0%) of the subjects made zero errors, and the remaining subjects got on average only about one question wrong. There is one interesting pattern, however. The averages are slightly higher in the process accountability conditions. This is entirely due to the question that probed the understanding of the accountability setup, suggesting that the process accountability setup was less intuitive or easy to understand than the outcome accountability setup. Note, however, that the total number of subjects who got this question wrong is low (10 out of the 120 process accountability subjects). In addition, they received corrective feedback immediately after the error, and their remaining data are not meaningfully different compared to the data of their counterparts.

Next, we checked subjects' motivation during the main task by classifying the trials according to inspected information. A trial was flagged if the subject failed to make a single inspection (conduct a single geological test) or made the decision before uncovering any information that would discriminate between the alternative they chose and the remaining two alternatives. A fully motivated subject who understood the task would have zero trials of this kind. This was the case for 80.8% of the subjects. Of the remaining ones, only two had 10% or more of their trials flagged and the differences among the experimental conditions are not meaningful.

Finally, we examined the average word count and response time on the question that asked the subjects to explain their decision-making approach. Because in (only) the process accountability condition the answer to this text would be shown to the second subject taking the

role of the manager (see *Experimental manipulations*), these measures can be thought of as a behavioral manipulation check of accountability type. Indeed, we observe robust differences in the expected direction. For example, the average number of words written is about 2.5 times higher among the process-accountable subjects.

Table 4.1: Condition means (standard deviations) for various measures (see section *Appendix 2* for definitions).

	Process acc. + Comp. env.	Outcome acc. + Comp. env.	Process acc. + Noncomp. env.	Outcome acc. + Noncomp. env.
Gender (% female)	53.3	65	57.6	55.9
Age (years)	20 (1.3)	20 (1.1)	20 (2.3)	20 (1.2)
Quiz (% errors)	6.4 (11.9)	2.8 (7)	7.9 (12.9)	3.1 (7.2)
Flagged trials (%)	0.3 (0.6)	0.6 (3)	0.5 (1.8)	0.3 (1.1)
Amount of search (%)	89.4 (15.6)	93.2 (11.1)	86.1 (17)	85.6 (15.3)
Validity search (%)	79.2 (35)	83.7 (28.5)	72.9 (37.7)	82.9 (27.8)
Time per test (avg.)	2.4 (0.9)	2.3 (0.9)	2.2 (1.1)	2.2 (0.8)
Accuracy, block 1 (%)	69.5 (8.8)	69.4 (7.6)	65.1 (7.1)	67.6 (6.3)
Accuracy, blocks 2-4 (%)	69.8 (6.3)	71.9 (5.1)	66.9 (6.6)	69.4 (4.1)
WADD (% search + choice)	62.4 (24.8)	69.6 (17.3)	56.3 (26.2)	51.5 (20.9)
WADD (% choice)	82.4 (10.5)	81.7 (8.5)	80.6 (9.9)	78.1 (11.4)
TTB (% search + choice)	67.1 (9.9)	65.5 (8.2)	71.4 (12.2)	75.9 (8.2)
Explanation (word count)	118.1 (65.9)	49.2 (33.1)	117.4 (70.4)	48 (29.6)
Explanation (seconds)	378.5 (253.2)	137.9 (89.5)	343.5 (234.3)	127.9 (87.5)
Optimism	4.2 (0.9)	4.2 (1)	4.1 (1)	3.9 (0.8)
Responsibility	4.2 (1.4)	3.7 (1.5)	3.8 (1.5)	3.8 (1.5)
Different approach	2.7 (1.8)	2.7 (1.6)	2.6 (1.7)	2.8 (1.6)
Rational thinking (avg.)	4.9 (0.6)	4.7 (0.8)	4.9 (0.8)	4.7 (0.8)

Predecisional behavior

Our setup allows us to examine three basic, independent aspects of predecisional behavior. First, subjects can differ in the *amount* of information they search for before deciding (for example, conducting more or fewer geological tests). Second, subjects can differ in the *order* they use in their search (for example, conducting the tests in order of their validity). Finally, subjects can differ in the amount of *time* they spend analyzing the information they acquired and deciding which alternative to choose. For each of these aspects, we first describe how we constructed the corresponding variable and then summarize the results. We articulate the overall takeaway of these analyses at the end of the section.

For the amount of search, we focused on the average proportion of tests conducted before making a choice. We call this measure *amount of search*. Row 5 in Table 4.1 reports the descriptive statistics and reveals that subjects in all four experimental conditions tended to conduct a rather comprehensive search. At the level of blocks, the measure is highly skewed, with about half of the observations corresponding to a fully comprehensive search

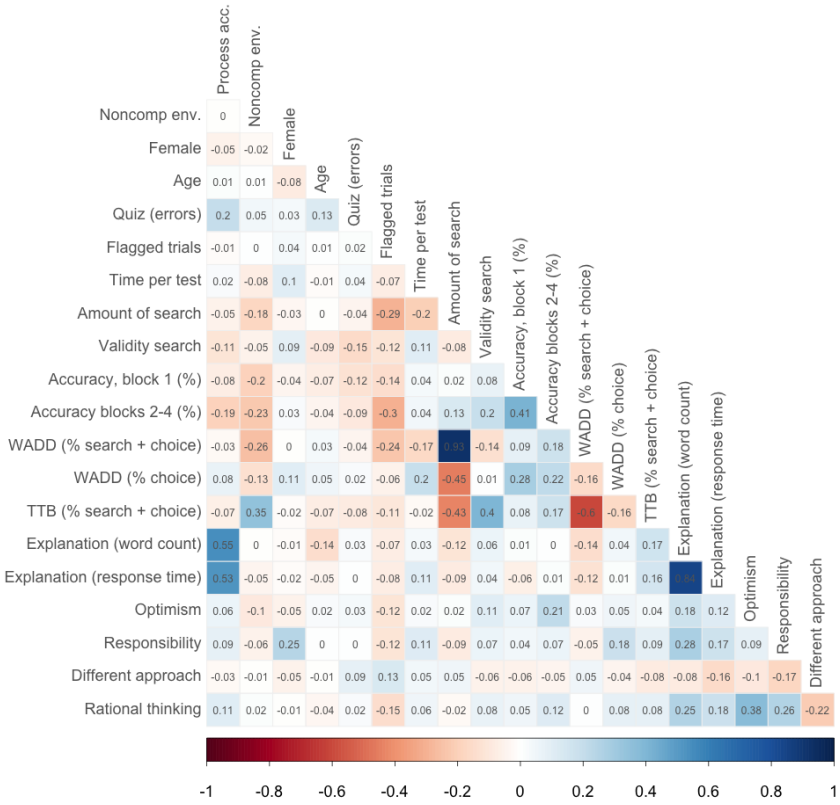


Figure 4.3: Correlations between the variables. For the definition of variables, see the main text or section *Appendix 2*. The correlations are Pearson's *r* correlations. (Color version of this figure is available in the electronic version of this document.)

(all tests conducted in all trials in a block; see Figure 4.7 in section *Appendix 3* for the full distribution). To simplify the modeling of this measure, we binned the values into four categories corresponding to always conducting all six tests (51.5% of observations), conducting at least five tests on average (19.4%), conducting at least four tests on average (15.9%), and the rest (13.3%). Afterwards, we fit a cumulative ordinal regression model with a probit link function (see Bürkner and Vuorre, 2019)[10]. The results are reported in column 1 of Table 4.2. The estimates (on the assumed latent scale) indicate the presence of main effects of environment and block, qualified by a two-way interaction between environment and block and a partial three-way interaction between environment, block, and accountability. To clarify these results, we

[10] This model assumes a latent normally distributed variable that gives rise to the observed categories based on to-be estimated cut-offs. Admittedly, this approach (categorizing the measure and fitting this sort of model) is not particularly satisfying.

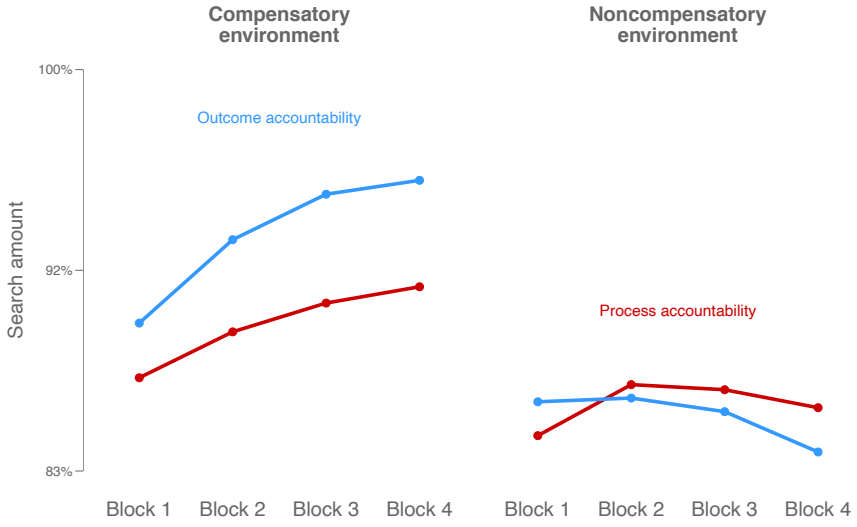
Table 4.2: Posterior means (standard deviations) for constant effects in each model.

Parameter	Amount of search	Validity search	Time per test	Accuracy	Rational thinking
Intercept			0.70 (0.02)***	69.09 (0.33)***	4.79 (0.05)***
Environment (non. vs. comp.)	-1.24 (0.40)***	-0.15 (0.24)	-0.07 (0.04)	-2.77 (0.67)***	0.03 (0.10)
Accountability (pro. vs. out.)	-0.51 (0.42)	-0.31 (0.24)	0.00 (0.04)	-2.04 (0.68)***	0.15 (0.09)
Block-A (2/3/4 vs. 1)	0.42 (0.11)***	1.07 (0.10)***	-0.53 (0.01)***	1.59 (0.42)***	
Block-B (3 vs. 2)	0.22 (0.14)	-0.06 (0.12)	-0.19 (0.02)***	0.94 (0.51)*	
Block-C (4 vs. 2/3)	-0.02 (0.12)	0.03 (0.10)	-0.18 (0.01)***	1.01 (0.42)**	
Env. x Acc.	1.13 (0.83)	-0.20 (0.48)	-0.03 (0.09)	-0.87 (1.37)	-0.01 (0.20)
Env. x Block-A	-0.57 (0.22)***	-0.14 (0.19)	0.04 (0.03)	0.39 (0.80)	
Env. x Block-B	-0.49 (0.28)*	-0.42 (0.23)*	0.04 (0.04)	-1.24 (0.98)	
Env. x Block-C	-0.64 (0.25)**	-0.26 (0.20)	-0.01 (0.03)	-1.27 (0.86)	
Acc. x Block-A	0.05 (0.22)	-0.38 (0.18)**	0.00 (0.03)	-1.06 (0.83)	
Acc. x Block-B	-0.25 (0.27)	-0.27 (0.24)	-0.04 (0.03)	0.52 (1.02)	
Acc. x Block-C	-0.16 (0.24)	0.06 (0.20)	-0.03 (0.03)	0.40 (0.86)	
Env. x Acc. x Block-A	1.00 (0.44)**	0.73 (0.37)**	0.09 (0.06)*	2.16 (1.64)	
Env. x Acc. x Block-B	0.18 (0.55)	0.45 (0.46)	-0.01 (0.07)	-1.86 (2.06)	
Env. x Acc. x Block-C	0.28 (0.47)	0.48 (0.40)	-0.07 (0.06)	-0.83 (1.71)	

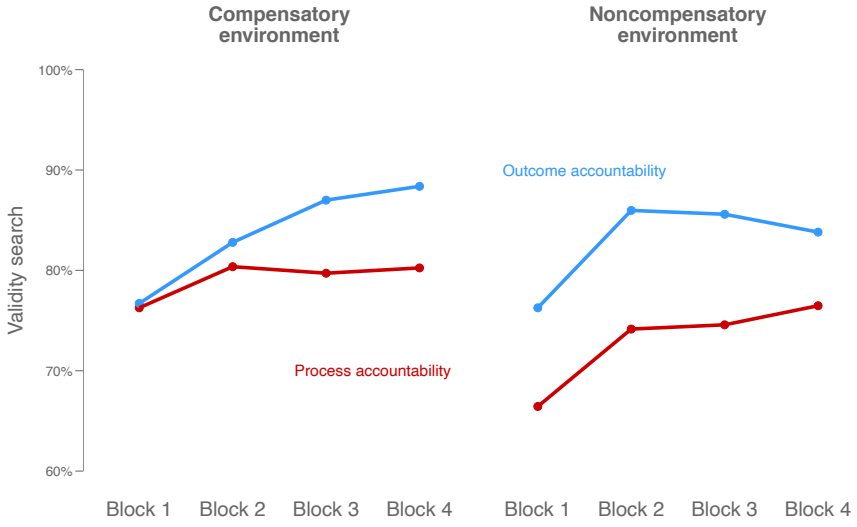
display the average amount of search (the original measure) in each experimental condition and each block in Figure 4.4a. Unsurprisingly, search is more comprehensive in the compensatory environment, where it increases across blocks in both accountability conditions but more so in the outcome accountability condition. The difference between the accountability conditions reverses in the noncompensatory environment, where process-accountable subjects on average conduct more tests after the first block. Given the small differences (note the range of the *y*-axis of Figure 4.4a) and our crude modeling approach, we should not put too much stock in the partial interaction between environment, block, and accountability. In any case, these results are clearly inconsistent with the idea that process-accountable individuals consider more information in their decision making than outcome-accountable ones.

Next, we analyzed the order of predecisional search. Although the order in which subjects conducted the geological tests is not important *per se*, it can provide insight into how subjects approached the task (for example, how systematic they were in their search process). We observed that in the large majority (79.7%) of trials, subjects would first conduct the geological test with the highest validity and then continue conducting tests in a descending order of validity[11]. We labeled this pattern *validity search* and constructed a measure of that same name that captures the proportion of trials in a given block in which a subject followed this pattern. This measure has a bimodal distribution, displayed in Figure 4.8 in section *Appendix 3*. For modeling purposes, we binned the values into four categories corresponding to following validity search in 100% of trials in a block (24.6% of observations), 90% or more (47.8%) trials in a block, 50% or more trials in a block (11.1%), or less than 50% (16.5%) of trials in a block. The modeling results (based on the same kind of ordinal regression model than the one used to model amount the of search) are summarized in column 2 of Table 4.2, and Figure 4.4b

[11]To be explicit, this measure does not require that all tests are conducted. It only requires that the tests that are conducted are conducted in a descending order of their validity.



(a) Amount of search



(b) Validity search

Figure 4.4: Amount and direction of search. The figure shows mean search amount (top) and mean validity-based search (bottom) across the four blocks, separately for each experimental condition. (Color version of this figure is available in the electronic version of this document.)

displays the average validity search (original measure) in every experimental condition and every block. We find that validity search increased in both decision environments after the introduction of feedback (cf. blocks 2-4 to block 1). Moreover, this increase was stronger under outcome accountability than under process accountability, particularly in the compensatory environment.

For the final aspect of predecisional behavior, we focused on decision time divided by the number of tests conducted (the division was done to make the measure independent from search amount). We call this measure *time per test*. Because we observed some large values which likely reflect contaminant processes, such as taking an unscheduled break during the task, we excluded 480 (1.3%) observations that fell three standard deviations away from the mean, with means and standard deviations computed separately for each subject and in log space. Because the data were skewed (see Figure 4.10 in section *Appendix 3*), we used the lognormal response distribution for modeling purposes. Descriptive statistics are reported in row 7 of Table 4.1, the modeling results are reported in column 3 of Table 4.2, and the condition averages for each block are displayed in Figure 4.9 in section *Appendix 3*. The main result is that time per test decreased monotonically as the task progressed (the decrease from the first to the last block being about 50%). The differences between the experimental conditions are overall minor.

What do these analyses of predecisional behavior jointly suggest about process and outcome accountability? We offer three tentative conclusions. First, there is no evidence for strong accountability differences on any of our measures. That is, we obtain no clear evidence for any main effect nor any interaction between accountability and environment. Second, to the extent that the data do indicate some potential differences, they are associated with the introduction of accuracy feedback in block 2. In particular, there is some evidence for a two-way interaction between accountability and the "Block-A" predictor on the validity measure search, and some evidence for a three-way interaction between accountability, environment, and the "Block-A" predictor on each of our three measures. Because the predictor "Block-A" captures the difference between blocks 2-4 (featuring accuracy feedback) and block 1 (no accuracy feedback), these results support the idea that it is important to study accountability effects in dynamic decision environments (featuring feedback) and not just static environments (without feedback), which were the focus of most of the previous work on process and outcome accountability. Finally, although it is not straightforward to interpret these two- and three-way interactions involving accountability, overall they paint a more favorable picture for *outcome* accountability (see Figures 4.4a and 4.4b), an interpretation that stands in contrast to previous work on the topic.

Accuracy

We used the percentage of correct decisions as a measure of accuracy. Condition means standard deviations are reported in Table 4.1, separately for block 1 (when no feedback was provided) and blocks 2-4 (when subjects did receive accuracy feedback after every decision). Modeling results are summarized in column 4 of Table 4.2 and reveal that subjects improved from block to block, performed better in the compensatory environment, and performed

better under outcome accountability. None of these effects is particularly large (about 1-3 percentage points) but the model is highly certain about their direction. For example, the model assigns a probability of 0.999 to the hypothesis that accuracy is on average higher under outcome accountability than under process accountability. In contrast to these main effects, the estimates of two-way and three-way interactions are associated with much more uncertainty. The crucial interaction between the type of environment and the type of accountability is in the predicted direction, but the uncertainty is too large to allow any conclusion about its direction.

Note that because the subjects did not receive any accuracy feedback in block 1, the results for this block are of lesser interest for the purpose of this analysis. Only blocks 2-4 directly correspond to our main research question, so in the next step we focus on these. As this is a Bayesian analysis, we can easily zero in on any comparison of interest and summarize it using the associated posterior probability distribution. To do this, we consider a hypothetical typical subject, examine what the model predicts about their accuracy in blocks 2-4 under all four experimental conditions, and summarize the results using the standard language of main effects and interactions. We find a main effect of environment type, accuracy being on average 2.7 percentage points (95% uncertainty interval: [0.9, 3.7], $P(\beta > 0) > 0.99$) higher in the compensatory environment. We also confirm the main effect of accountability, accuracy being on average 2.3 percentage points (95% uncertainty interval: [1.3, 4.1], $P(\beta > 0) > 0.99$) higher under outcome accountability. Finally, the difference between outcome and process accountability is indeed higher in the noncompensatory environment (for about 0.4 percentage points), but this estimate is small and too uncertain (95% uncertainty interval: [-2.5, 3.1]), $P(\beta > 0) \approx 0.61$) to allow any conclusion about the direction of the interaction.

To obtain an even more fine-grained (and less model-based) comparison of process and outcome accountability in the two decision environments, we computed a *moving average* using a window of 40 trials (the length of one block). Thus the first observation for each subject is computed by averaging over trials 1 - 40, the second one over trials 2 - 41, etc. We display the results in Figure 4.5. The lines correspond to the mean values for each experimental condition and the shaded areas capture the corresponding 95% uncertainty intervals (the latter were obtained using a bootstrapping procedure with 1000 iterations). In the compensatory environment, we observe that the mean accuracy is essentially the same for process and outcome accountability in block 1 (trial 40 in the figure). But outcome-accountable subjects respond better to the introduction of feedback in block 2 and perform better throughout the rest of the task. In the noncompensatory environment, the mean accuracy is consistently higher under outcome accountability and the difference stays relatively constant throughout the task.

Overall—in contrast to the dominant finding in the accountability literature but in agreement with our preliminary study—the results indicate a small but reliable performance edge for outcome accountability in both the compensatory and the noncompensatory decision environment.

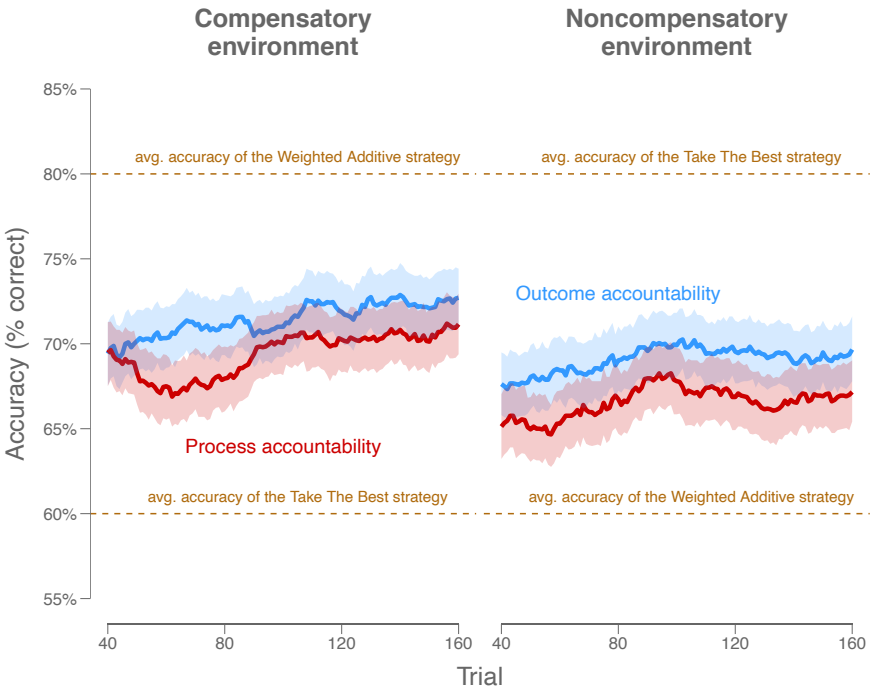


Figure 4.5: Accuracy (% of correct decisions) throughout the task. The measure depicted on the *y*-axis is the rolling average, computed using a window of 40 trials (length of one block). The lines denote condition averages, and the shaded areas display 95% uncertainty intervals obtained using a bootstrapping procedure (method: percentile-based, number of iterations: 1000). (Color version of this figure is available in the electronic version of this document.)

Decision strategy

Differences in predecisional behavior and accuracy could be driven by differences in the underlying decision strategies that subjects employed. To investigate this possibility rigorously, we combined search data and choice data. Any given trial was considered to be compatible with a given strategy if and only if (i) the minimal number of geological tests required by the strategy was conducted (*criterion of minimal search*), and (ii) the choice matched the strategy prediction (*choice criterion*). For example, if a subject conducted all geological tests but did not choose the oil-drilling site favored by the Weighted Additive strategy, the trial was considered incompatible with the Weighted Additive strategy. Similarly, if the choice matched the prediction of Take the Best but the most valid geological test was not conducted, the trial was considered to be incompatible with the Take The Best strategy. Etc.

On the basis of this compatibility measure, we classified each subject as either a Weighted Additive user, a Take The Best user, or an Equal Weights user[12]. We did this twice, once

[12] The Equal Weights strategy is a special case of the Weighted Additive strategy with all weights set to the same

for the first block and once for the final block. The former classification captures decision strategy adoption in the absence of accuracy feedback, and the latter classification captures strategy adoption after the provision of considerable accuracy feedback. In the first block, the results show a relatively even split between users of the Weighted Additive strategy and users of the Take The Best strategy, with the latter somewhat more common (52.6%-61.8%, depending on the exact condition[13]). The differences between experimental conditions are minor. In contrast, clear differences emerge in the final block. In particular, most subjects were classified as users of the most adaptive strategy in their decision environment. In the compensatory environment, users of the Weighted Additive strategy predominated (60.7% in the process accountability condition, 72.9% in the outcome accountability condition). And in the noncompensatory environment, users of the Take The Best strategy predominated (76.8% in the process accountability condition, 87.5% in the outcome accountability condition)[14]. These percentages also reveal that in both environments, outcome-accountable subjects were more likely to adopt the most adaptive strategy.

To obtain an alternative and more fine-grained look into these data, we conducted an analysis based on a moving average. This analysis is analoguous to the one reported for accuracy in Figure 4.5. The only difference is that we used compatibility with the most adaptive strategy as the dependent variable instead of accuracy. Thus for the compensatory (noncompensatory) environment, the measure expresses how often subjects behaved in accordance with the Weighted Additive (Take The Best) strategy. The results are summarized in Figure 4.6. The figure reveals a clear pattern of adaptation in all four experimental conditions, with the average compatibility increasing throughout the task. In both decision environments, outcome-accountable subjects are more likely to behave in accordance with the most adaptive strategy. The results for the compensatory environment are especially interesting, as they clearly contradict the idea that process-accountable subjects would be more likely to decide in accordance with the Weighted Additive strategy.

In light of previous research on this topic, one curious feature of these results is the high apparent prevalence of Take The Best users. In studies of this sort that lack explicit search costs and withhold accuracy feedback, the Weighted Additive strategy is typically favored by the majority of subjects. In contrast, in the first block of the current study (no explicit search costs, no accuracy feedback) the Take The Best users were in fact slightly more common. To investigate this discrepancy, we conducted an additional analysis in which we relaxed the criterion of minimal search. That is, a trial was considered compatible with a given strategy as long as the strategy prediction *given the information the subject actually acquired* matched the subject's choice. In essence, this analysis asks about the *choice mechanism* that subjects employed—did they follow the compensatory approach of the Weighted Additive strategy or the lexicographic, noncompensatory approach of the Take The Best strategy? The results of this analysis provide an important qualification to our primary analysis. In line with previous literature, in the first block most subjects (75.4%-83.9%, depending on the condition) were classified as users of the Weighted Additive strategy. In the compensatory environment, the

positive constant.

[13] This is based only on the subjects for which one strategy provided a uniquely best fit (96.2% of subjects). For the remaining 3.8%, the Weighted Additive and Take The Best strategies had the exact same fit.

[14] As for block 1, there was a small group of subjects (2.9%) for which the Weighted Additive and Take The Best strategies provided the exact same fit and are therefore not included in this classification.

proportion of Weighted Additive users remained high in the final block (88.1% in both the process and the outcome accountability condition). In contrast, in the noncompensatory environment the proportion of Weighted Additive users decreased markedly (to 50.9% in the process accountability condition and 45.3% in the outcome accountability condition) — but nevertheless stayed relatively high. The contrast between this analysis and the primary analysis suggests that a substantial minority of subjects paired the choice mechanism of the Weighted Additive strategy with a more limited search behavior characteristic of the Take The Best strategy.

Postexperimental questionnaire

Rows 16 - 21 in Table 4.1 report summary statistics for the measures collected after the main task was completed. We have already discussed the results for the *explain-your-approach* question in the section *Preliminary analyses*. For the remaining ones, we note that the majority of subjects expressed optimism about their evaluation (78.4%), felt some responsibility towards the second subject (64.0%), and indicated they would use a similar approach if decision outcomes *directly* affected themselves (65.3%)[15]. None of these measures correlates strongly with other measures or displays meaningful differences between the experimental conditions. Finally, the postexperimental questionnaire included the scale of situation-specific rational thinking style (Novak and Hoffman, 2009). Figure 4.3 shows that the average on this scale correlates positively with accuracy after the first block, although the correlation is low (Pearson's $r = 0.12$, 95% classical uncertainty interval: [-0.01, 0.24]). In line with de Langhe et al. (2011), we find that the scores are on average somewhat higher among process-accountable subjects. However, in the current study the difference is about half the magnitude of the one reported by de Langhe et al. (2011), whether we consider the raw effect size expressed as a percentage of maximum possible difference (7.0% vs. 3.1%) or a standardized effect size (Cohen's d: 0.48 vs. 0.21). In summary, in the current study the difference in rational thinking style between process and outcome accountability is relatively small and also consistent with an absence of an effect, although the model (summarized in column 5 of Table 4.2) does assign a probability of almost 0.95 to the hypothesis that the true average score on this measure is higher under process accountability.

4.4 Discussion

In the current research we investigated decision accuracy and cognitive processing of decision makers under either process or outcome accountability. Most previous studies in this stream of literature focused on tasks in which decision makers possess (or are presented with) some relevant background knowledge but do not receive any outcome feedback on the accuracy of their decisions. In contrast, we utilized a dynamic decision-making task in which decision makers face the need to combine reliable historical information and outcome feedback. Our study featured two decision environments, a compensatory one and a noncompensatory one.

[15] These percentages are based on a dichotomization of responses based on the midpoint of the scale, i.e. 3.5.

Whereas optimal accuracy in the compensatory environment called for the use of a compensatory, information-intensive strategy, it was facilitated by the use of a noncompensatory, information-frugal strategy in the noncompensatory environment. Results of simulations involving real-world datasets suggest that environments of the latter kind may be common, but they have so far not been investigated in the accountability literature. Importantly, previous work suggested that process accountability (relative to outcome accountability) facilitates a decision-making approach that involves an increased search for information and a compensatory integration of acquired information. This kind of approach lends itself well to a compensatory environment but could prove maladaptive in a noncompensatory environment. Testing the relative effects of process and outcome accountability on decision accuracy in compensatory and noncompensatory environments was the primary objective of the current study.

We conducted a 2 × 2 between-subjects laboratory experiment and tracked subjects' decision accuracy throughout the task. As predicted, in the noncompensatory environment the average accuracy was reliably higher under outcome accountability than under process accountability. But the results also yielded a surprise: outcome-accountable subjects outperformed process-accountable ones in the compensatory environment as well, and the magnitude of the effect did not differ meaningfully compared to the noncompensatory environment. Overall, although the predicted interaction effect did not materialize, these results provide an important contribution to the literature. Comparable laboratory research on process and outcome accountability previously found either higher performance under process accountability or, in a few cases, similar levels of performance under process and outcome accountability. In contrast, our results suggest that in dynamic decision environments that feature outcome feedback, outcome accountability can be preferable to process accountability.

The second objective of the current research was to investigate cognitive processing under process and outcome accountability. To pursue this objective, we relied on both process-tracing (Schulte-Mecklenbeck et al., 2017) and self-report (Novak and Hoffman, 2009). For the former, we used a presentation format in which the decision information was initially hidden, allowing us to track how much information the subjects obtained, in what order, and how long they (presumably) spent analyzing the acquired information. For the latter, we used the situation-specific rational thinking style questionnaire that was previously used in explaining the superior performance of process-accountable individuals (de Langhe et al., 2011). Starting with the self-report data, we found that the average rational thinking score was higher under process accountability, in line with de Langhe et al. (2011). However, our data did not provide convincing evidence for this effect, and its magnitude was only about 50% of the magnitude reported by de Langhe et al. (2011). A closer look revealed that this difference in magnitude stemmed from the relatively high scores of outcome-accountable subjects in the current study—whereas the scores of process-accountable subjects were very similar to those of their counterparts in de Langhe et al. (2011). While these self-report results provide mixed evidence, the process-tracing results paint a clearer picture. None of the three measures we employed yielded any evidence of more systematic, deliberate, and comprehensive predecisional behavior under process accountability. In particular, the amount of information acquired before a decision was taken, the extent to which search was systematic, and the time spent analyzing the acquired information were all similar under process and outcome

accountability.

Altogether, the results of the current study challenge the view of process and outcome accountability supported by the existing academic literature. This discrepancy may be due to the unique challenges posed by decision tasks that combined static, historical information and continuous outcome feedback. But it is worthwhile asking what other characteristics of the current study could have contributed to it. Perhaps the most salient such characteristic concerns the specific form our accountability manipulations took. In most previous studies, the evaluation of the accountable individuals was done by the researcher conducting the study (or their associate). In contrast, the evaluation in the current study was done by another subject in the study. Although this difference (and the corresponding variation in the degree of informational asymmetry between the accountable individual and the evaluator) could potentially be meaningful and an interesting target for future research, it is unlikely to have played a major role here. That is because the preliminary study (see *Introduction*) used the more typical setup (evaluation done by the researcher) and found overall similar results. Another salient design feature of the current study is the use of a presentation format with initially-hidden decision information. Although this design choice allowed us to track the (maximum) amount of information subjects relied on when making a decision, the requirement to explicitly request decision-relevant information possibly facilitated a more deliberate approach (Söllner et al., 2013) and thus diluted the differences between process and outcome accountability that would otherwise have manifested. For example, this design choice could have contributed to the weaker effect on the rational thinking style questionnaire. It is not clear, however, how this design choice could have contributed to the higher decision accuracy under outcome accountability.[16]

Interestingly, although our results go against the view of process and outcome accountability supported by *laboratory* research, they do resonate with the results of a recent *field* experiment. Chang et al. (2017) tracked the long-term performance of 1850 forecasters and found that outcome-accountable forecasters outperformed process-accountable ones and that the difference grew over time. It is important to note, however, that the evaluation of process-accountable subjects in Chang et al. (2017) centered on a specific set of forecasting guidelines, whereas in our study no explicit decision-making guidelines were communicated to the subjects. This is important because laboratory research suggests that the benefits of process accountability are maximized when the evaluation is relatively open-ended and not based on guidelines communicated in advance (Patil et al., 2014, 2017). Our results therefore suggest an important limitation of process accountability even when implemented in the form that previous literature suggests is most beneficial.

Future research could shed further light on the relationship between the properties of a decision environment, the properties of feedback, and the relative effects of process and outcome accountability. For example, feedback can take many forms, it can be more or less timely, more or less complete, and more or less informative. It can also vary in the ease with which it can be communicated to other parties. The distinction between *kind* and *wicked* learning environments (Hogarth et al., 2015) is relevant and could inform future research. Another important characteristic of business reality is that market conditions sometimes change and

[16] Siegel-Jacobs and Yates (1996) also used an information board approach and found higher judgment accuracy under process accountability than under outcome accountability.

leave a previously successful strategy inappropriate for dealing with the new environment. Future research could investigate if the relative advantage of outcome accountability that we observed in this study would be even more pronounced under these conditions.

In conclusion, the current research suggests that the existing accountability literature overemphasizes the relative benefits of process accountability and that, at least under certain conditions, outcome accountability supports more adaptive decision making. At the same time, the above observations make it clear that we should be wary of generalizing the results of the current study, at least until a better understanding of the underlying mechanism(s) is developed. In order to inform the complex reality of business practice, accountability research needs to develop a more precise, complete, and genuinely predictive framework for the relative relative effects of process and outcome accountability.

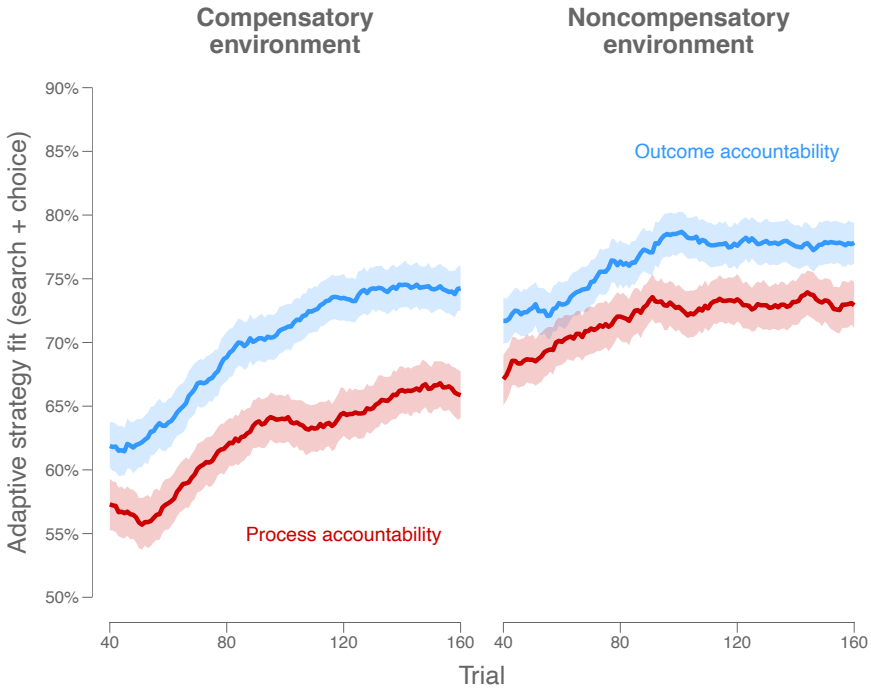


Figure 4.6: Compatibility with the adaptive strategy throughout the task. The adaptive strategy was Weighted Additive in the compensatory environment and Take The Best in the noncompensatory environment. A given trial is compatible with a given strategy iff (i) the decision matches the strategy prediction and (ii) at least the minimal number of geological tests required by the strategy was conducted. The measure depicted on the y-axis is the moving average, computed using a window of 40 trials (length of one block). The lines denote condition averages, and the shaded areas display 95% uncertainty intervals obtained using a bootstrapping procedure (method: percentile-based, number of iterations: 1000). (Color version of this figure is available in the electronic version of this document.)

4.5 Appendix 1

This appendix provides information on the exact decision trials used in the study.

Table 4.3 reproduces the (unique) trials used in the compensatory decision environment. The first column identifies each trial, the following columns correspond to the three alternatives defined by six cues. For example, columns 2 - 7 corrrespond to the values of decision alternative 1 on cues 1 - 6 (hence the headers C1 - C6). A value of 1 corresponds to a positive test result and a value of 0 corresponds to a negative test result.

Table 4.4 reproduces the (unique) trials used in the noncompensatory decision environment.

Table 4.3: Trials used in the compensatory environment.

ID	Alternative 1						Alternative 2						Alternative 3					
	C1	C2	C3	C4	C5	C6	C1	C2	C3	C4	C5	C6	C1	C2	C3	C4	C5	C6
1	0	1	1	0	0	1	0	1	1	0	0	0	1	0	0	0	0	1
2	0	1	0	1	1	0	0	1	0	0	1	0	1	0	0	1	0	0
3	1	1	1	0	1	1	1	0	1	0	0	0	1	0	1	1	1	1
4	1	0	1	1	0	1	0	1	0	1	0	0	1	1	1	0	0	0
5	1	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	1
6	1	1	0	0	1	1	0	0	1	0	0	0	0	1	1	1	1	1
7	0	0	0	1	1	1	0	0	1	1	0	0	0	0	0	0	1	1
8	1	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1
9	0	1	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	1
10	0	1	0	1	1	0	0	1	0	0	0	0	0	1	0	1	0	1
11	1	0	1	1	1	1	1	1	0	1	1	0	1	1	0	0	0	0
12	1	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1
13	1	1	0	0	1	1	0	0	0	1	1	0	0	1	1	1	0	0
14	1	1	1	0	0	0	0	0	1	0	1	1	1	1	0	0	1	1
15	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0
16	1	1	0	1	1	0	0	0	1	0	0	1	1	0	0	1	1	1
17	0	1	0	0	0	1	1	1	0	0	1	0	1	0	1	0	1	1
18	0	0	1	1	1	0	0	1	1	0	0	1	1	0	0	0	0	1
19	1	1	1	0	1	1	1	0	1	0	0	0	1	1	1	0	1	0
20	0	1	0	1	1	0	0	1	0	1	0	1	1	0	0	0	1	0
21	1	0	1	0	0	0	1	1	0	1	0	0	1	0	1	0	1	1
22	1	0	1	0	1	0	0	1	1	1	1	0	1	1	0	0	0	1
23	1	1	1	1	0	0	0	0	1	0	0	1	1	0	0	0	1	0
24	0	0	0	0	1	1	0	0	1	1	1	0	0	1	1	0	0	1
25	0	1	1	1	0	1	0	0	0	1	1	0	0	0	0	0	1	1
26	1	1	0	0	0	1	0	1	1	1	1	0	0	1	1	0	0	1
27	0	1	1	1	1	1	0	0	0	1	0	0	1	1	0	1	0	0
28	1	1	0	1	0	1	0	0	0	0	1	1	0	1	1	1	0	1
29	1	0	1	1	1	0	0	0	0	1	0	1	0	1	1	0	0	0
30	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	1	0
31	0	1	0	1	1	0	0	0	0	0	0	1	1	0	0	1	0	0
32	1	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0
33	1	1	1	0	1	1	0	0	0	0	0	0	1	0	0	1	1	1
34	1	1	1	1	0	1	0	1	1	0	1	0	0	0	1	1	1	0
35	0	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1
36	1	0	0	1	1	0	0	0	0	1	0	1	0	1	1	0	0	0
37	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0	0
38	1	0	1	1	1	1	1	1	0	0	0	0	1	0	0	1	1	1
39	1	1	1	1	0	1	0	1	0	1	0	0	0	1	1	1	0	0
40	1	1	1	0	0	0	0	1	1	0	1	1	0	0	1	0	0	0

Table 4.4: Trials used in the noncompensatory environment.

ID	Alternative 1						Alternative 2						Alternative 3					
	C1	C2	C3	C4	C5	C6	C1	C2	C3	C4	C5	C6	C1	C2	C3	C4	C5	C6
1	0	1	1	0	0	1	0	1	1	0	0	0	1	0	0	0	0	1
2	0	1	0	1	1	0	0	1	0	0	1	0	1	0	0	1	0	0
3	1	1	1	0	1	1	1	0	1	0	0	0	1	0	1	1	1	1
4	1	0	1	1	0	1	0	1	0	1	0	0	1	1	1	0	0	0
5	1	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	1
6	1	1	0	0	1	1	0	0	1	0	0	0	0	1	1	1	1	1
7	0	0	0	1	1	1	0	0	1	1	0	0	0	0	0	0	1	1
8	1	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1
9	0	1	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	1
10	0	1	0	1	1	0	0	1	0	0	0	0	0	1	0	1	0	1
11	1	0	1	1	1	1	1	1	0	1	1	0	1	1	0	0	0	0
12	1	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1
13	1	1	0	0	1	1	0	0	0	1	1	0	0	1	1	1	0	0
14	1	1	1	0	0	0	0	0	1	0	1	1	1	1	0	0	1	1
15	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0
16	1	1	0	1	1	0	0	0	1	0	0	1	1	0	0	1	1	1
17	0	1	0	0	0	1	1	1	0	0	1	0	1	0	1	0	1	1
18	0	0	1	1	1	0	0	1	1	0	0	1	1	0	0	0	0	1
19	1	1	1	0	1	1	1	0	1	0	0	0	1	1	1	0	1	0
20	0	1	0	1	1	0	0	1	0	1	0	1	1	0	0	0	1	0
21	1	0	1	0	0	0	1	1	0	1	0	0	1	0	1	0	1	1
22	1	0	1	0	1	0	0	1	1	1	1	0	1	1	0	0	0	1
23	1	1	1	1	0	0	0	0	1	0	0	1	1	0	0	0	1	0
24	0	0	0	0	1	1	0	0	1	1	1	0	0	1	1	0	0	1
25	0	1	1	1	0	1	0	0	0	1	1	0	0	0	0	0	1	1
26	1	1	0	0	0	1	0	1	1	1	1	0	0	1	1	0	0	1
27	0	1	1	1	1	1	0	0	0	1	0	0	1	1	0	1	0	0
28	1	1	0	1	0	1	0	0	0	0	1	1	0	1	1	1	0	1
29	1	0	1	1	1	0	0	0	0	1	0	1	0	1	1	0	0	0
30	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	1	0
31	0	1	0	1	1	0	0	0	0	0	0	1	1	0	0	1	0	0
32	1	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0
33	1	1	1	0	1	1	0	0	0	0	0	0	1	0	0	1	1	1
34	1	1	1	1	0	1	0	1	1	0	1	0	0	0	1	1	1	0
35	0	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1
36	1	0	0	1	1	0	0	0	0	1	0	1	0	1	1	0	0	0
37	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0	0
38	1	0	1	1	1	1	1	1	0	0	0	0	1	0	0	1	1	1
39	1	1	1	1	0	1	0	1	0	1	0	0	0	1	1	1	0	0
40	1	1	1	0	0	0	0	1	1	0	1	1	0	0	1	0	0	0

4.6 Appendix 2

Table 4.4: Variable definitions.

Label	Definition
Quiz (% errors)	Percentage of errors made on the 6-question comprehension quiz administered before the main task.
Flagged trials (%)	Percentage of trials in the main task that were flagged because the subject (a) failed to conduct any geological tests (meaning that the decision on that trial was essentially random) or (b) failed to uncover any information that would discriminate between the oil-drilling site eventually chosen and one or both of the remaining oil-drilling sites.
Amount of search (%)	Average percentage of geological tests conducted per trial.
Validity search (%)	Percentage of trials in which validity-based search order was employed. Validity-based search order means that the most valid test was conducted first, and any further test that was conducted was always the test with the highest validity among the remaining tests.
Time per test (avg.)	Average time per test, where "time per test" corresponds to decision time (number of seconds elapsed from the moment a trial started to the moment the decision was made) divided by the number of tests conducted in that trial. Outliers were excluded at the trial-level, see the main text.
Accuracy, block 1 (%)	Percentage of correct decisions made in the first block of the main task.
Accuracy, blocks 2-4 (%)	Percentage of correct decisions made in blocks 2–4 of the main task.
WADD (% search + choice)	Percentage of trials compatible with the Weighted Additive decision strategy in terms of both search (was the minimum number of tests needed to carry out the strategy conducted?) and choice (did the decision match the strategy prediction?).
WADD (% choice)	Percentage of trials compatible with the Weighted Additive decision strategy in terms of choice (did the decision match the prediction of the strategy *given* the information uncovered by the subject?).

Label	Definition
TTB (% search + choice)	Percentage of trials compatible with the Take The Best strategy in terms of both search (was the minimum number of tests needed to carry out the strategy conducted?) and choice (did the decision match the strategy prediction?). Because the Take The Best requires only a rather frugal search, this measure was almost the same as TTB (% choice).
Explanation (word count)	Number of words written in response to the *explain-your-approach* question.
Explanation (seconds)	Time (in seconds) taken to submit the response to the *explain-your-approach* question.
Optimism	Answer to the question "Do you believe you will be evaluated favorably by the manager?". The question was part of the post-experimental questionnaire, the response scale was a 1–6 Likert scale with anchors 1 = No, not at all and 6 = Yes, for sure.
Responsibility	Answer to the question "Have you felt any responsibility for how your decisions would affect the second participant (the manager)?". The question was part of the postexperimental questionnaire, the response scale was a 1–6 Likert scale with anchors 1 = No, not at all and 6 = Yes, for sure.
Different approach	Answer to the question "Would your decision making [*sic*] approach be different if decision outcomes directly affected you (as opposed to the manager)?". The question was part of the postexperimental questionnaire, the response scale was a 1–6 Likert scale with anchors 1 = No, it would be the same and 6 = Yes, it would be very different.
Rational thinking (avg.)	Average of the ten items of the rational subscale of the situation-specific thinking style questionnaire (Novak and Hoffman, 2009). The response scale was a 1–6 Likert scale with anchors 1 = Completely disagree and 6 = Completely agree.

4.7 Appendix 3

This appendix reports some additional results and comments on the results reported in the main text. Its structure matches the structure of the *Results* section.

Preliminary analyses

The main text mentions ten subjects (all in the process accountability condition) who incorrectly answered the comprehension quiz question probing their understanding of the accountability setup. The main text also mentions two subjects who had an unusually large number of "trials" flagged[17]. To check if it makes a difference, we re-ran every model reported in Table 4.2 after excluding these 12 subjects. The results are reported in Table 4.5.

Table 4.5: Posterior means (standard deviations) for constant effects in each model.

Parameter	Search amount	Validity search	Time per test	Accuracy	Rational thinking
Intercept			0.71 (0.02)***	69.26 (0.34)***	4.79 (0.05)***
Environment (non. vs. comp.)	-1.35 (0.43)***	-0.27 (0.24)	-0.07 (0.05)	-2.68 (0.68)***	0.03 (0.10)
Accountability (pro. vs. out.)	-0.67 (0.41)	-0.21 (0.23)	0.01 (0.04)	-1.91 (0.69)***	0.15 (0.10)
Block-A (2/3/4 vs. 1)	0.39 (0.11)***	1.09 (0.11)***	-0.52 (0.01)***	1.51 (0.41)***	
Block-B (3 vs. 2)	0.19 (0.15)	-0.06 (0.12)	-0.19 (0.02)***	0.89 (0.50)*	
Block-C (4 vs. 2/3)	-0.01 (0.12)	0.01 (0.10)	-0.18 (0.02)***	1.13 (0.43)**	
Env. x Acc.	1.23 (0.90)	-0.30 (0.50)	-0.02 (0.09)	-0.33 (1.36)	0.01 (0.20)
Env. x Block-A	-0.66 (0.23)***	-0.05 (0.19)	0.02 (0.03)	0.10 (0.82)	
Env. x Block-B	-0.43 (0.29)	-0.28 (0.25)	0.03 (0.04)	-0.91 (1.01)	
Env. x Block-C	-0.57 (0.25)**	-0.24 (0.21)	-0.01 (0.03)	-1.28 (0.90)	
Acc. x Block-A	-0.13 (0.23)	-0.39 (0.19)**	0.02 (0.03)	-1.41 (0.84)*	
Acc. x Block-B	-0.34 (0.28)	-0.23 (0.24)	-0.04 (0.04)	0.20 (1.01)	
Acc. x Block-C	-0.16 (0.24)	0.10 (0.22)	-0.04 (0.03)	0.32 (0.86)	
Env. x Acc. x Block-A	1.07 (0.46)**	0.87 (0.39)**	0.05 (0.06)	1.92 (1.67)	
Env. x Acc. x Block-B	0.36 (0.58)	0.65 (0.49)	-0.02 (0.07)	-0.68 (2.13)	
Env. x Acc. x Block-C	0.47 (0.51)	0.40 (0.42)	-0.07 (0.06)	-0.23 (1.74)	

Predecisional behavior

Amount of search. The main text mentions the skewed distribution of this measure. Figure 4.7 displays it, along with the cut-offs used when transforming the variable for modeling purposes (dashed red lines). These cut-offs correspond to always conducting all six tests in each trial of the block, conducting on average five or more tests, conducting on average four or more tests, and the rest (the pattern does not continue because there would be very few observations in the following bins). In the main text, we used Bayesian estimation to model the resulting measure. Column 1 in Table 4.6 reports the results of an equivalent model when maximum likelihood estimation is used instead.

[17] I am but a lonely footnote.

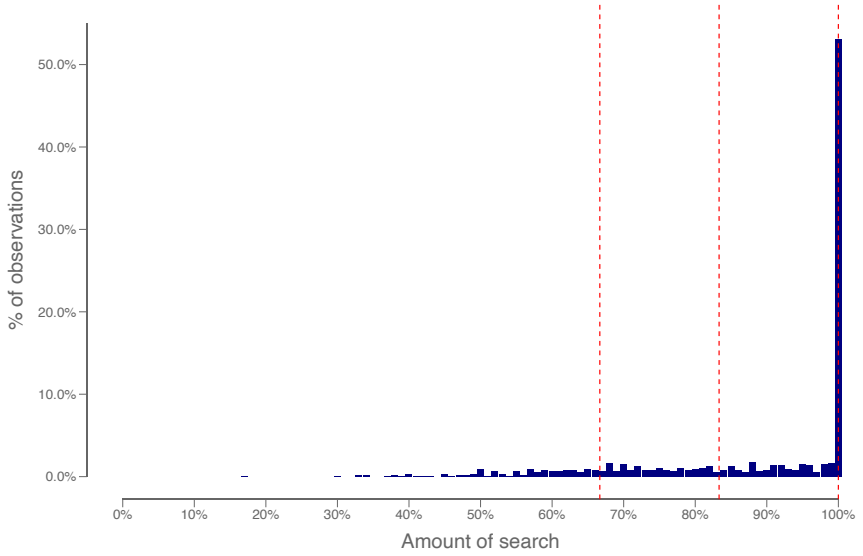


Figure 4.7: Distribution of the search amount measure. (Color version of this figure is available in the electronic version of this document.)

Validity search. The main text mentions the bimodal distribution of this measure. Figure 4.8 displays it, along with the cut-offs used when transforming the variable for modeling purposes (dashed red lines). These cut-offs correspond to using validity-based search in exactly 100% / 90% or more / 50% or more / less than 50% trials in a block. In the main text, we used Bayesian estimation to model the resulting measure. Column 2 in Table 4.6 reports the results of an equivalent model when maximum likelihood estimation is used instead.

Time per test. Figure 4.9 shows the mean time per test in every block and every experimental condition, and Figure 4.10 shows the distribution of the measure. In the main text, we used Bayesian estimation to model the measure. Column 3 in Table 4.6 reports the results of an equivalent model when maximum likelihood estimation is used instead.

Accuracy

In the main text, we used Bayesian estimation when modeling accuracy. Column 4 in Table 4.6 reports the results of the modeling analysis when maximum likelihood estimation is used instead.

Postexperimental questionnaire

In the main text, we used Bayesian estimation to analyze the effects of environment and accountability on rational thinking, measured using the scale of Novak and Hoffman (2009).

Table 4.6: Classical estimates (standard errors).

Parameter	Search amount	Validity search	Time per test	Accuracy	Rational thinking
Intercept			0.70 (0.02)***	69.10 (0.34)***	4.79 (0.05)***
Environment (non. vs. comp.)	-1.26 (0.38)***	-0.17 (0.22)	-0.07 (0.04)	-2.76 (0.68)***	0.03 (0.10)
Accountability (pro. vs. out.)	-0.53 (0.37)	-0.30 (0.22)	0.01 (0.04)	-2.02 (0.68)***	0.15 (0.10)
Block-A (2/3/4 vs. 1)	0.41 (0.11)***	1.05 (0.10)***	-0.53 (0.01)***	1.59 (0.41)***	
Block-B (3 vs. 2)	0.21 (0.14)	-0.05 (0.12)	-0.19 (0.02)***	0.94 (0.50)*	
Block-C (4 vs. 2/3)	-0.03 (0.12)	0.03 (0.10)	-0.18 (0.02)***	1.01 (0.43)**	
Env. x Acc.	1.17 (0.74)	-0.19 (0.44)	-0.04 (0.09)	-0.93 (1.36)	-0.01 (0.19)
Env. x Block-A	-0.56 (0.22)**	-0.13 (0.19)	0.04 (0.03)	0.40 (0.82)	
Env. x Block-B	-0.46 (0.27)*	-0.41 (0.23)*	0.04 (0.04)	-1.25 (1.00)	
Env. x Block-C	-0.61 (0.24)**	-0.26 (0.20)	-0.01 (0.03)	-1.28 (0.86)	
Acc. x Block-A	0.05 (0.22)	-0.37 (0.19)**	0.00 (0.03)	-1.07 (0.82)	
Acc. x Block-B	-0.23 (0.27)	-0.26 (0.23)	-0.04 (0.04)	0.54 (1.00)	
Acc. x Block-C	-0.15 (0.24)	0.06 (0.20)	-0.04 (0.03)	0.39 (0.86)	
Env. x Acc. x Block-A	0.96 (0.44)**	0.70 (0.37)*	0.10 (0.06)*	2.16 (1.63)	
Env. x Acc. x Block-B	0.15 (0.55)	0.46 (0.46)	-0.01 (0.07)	-1.84 (2.00)	
Env. x Acc. x Block-C	0.25 (0.47)	0.48 (0.40)	-0.07 (0.06)	-0.82 (1.73)	

That analysis is reported in column 5 of Table 4.2. Column 5 of Table 4.6 reports the results for an equivalent model fit using maximum likelihood estimation.

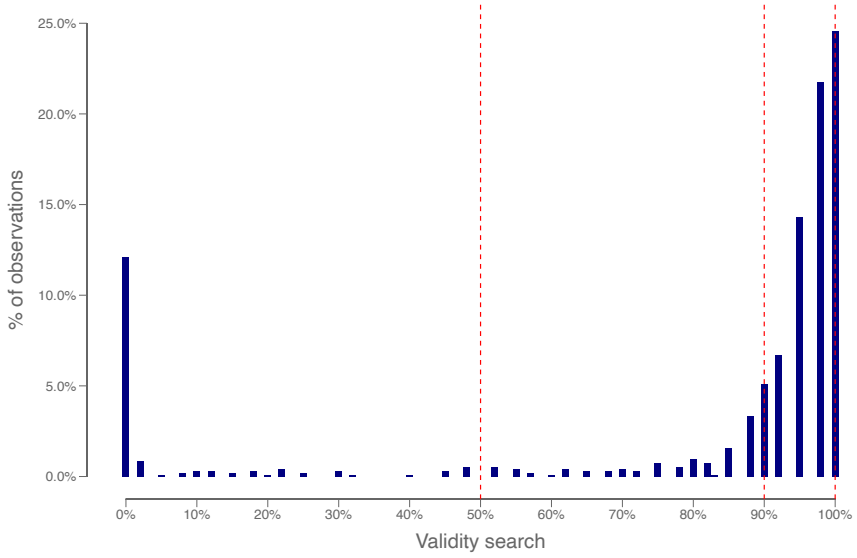


Figure 4.8: Distribution of the validity search measure. (Color version of this figure is available in the electronic version of this document.)

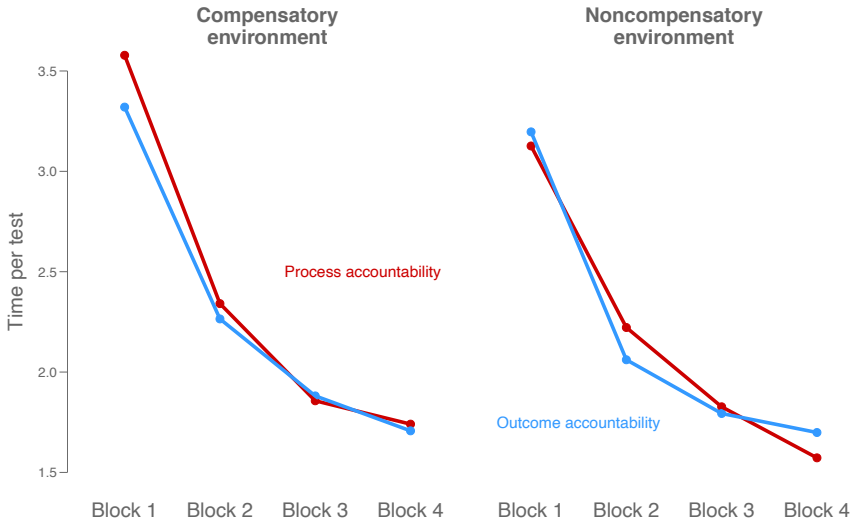


Figure 4.9: Average time per test in each block and each experimental condition. (Color version of this figure is available in the electronic version of this document.)

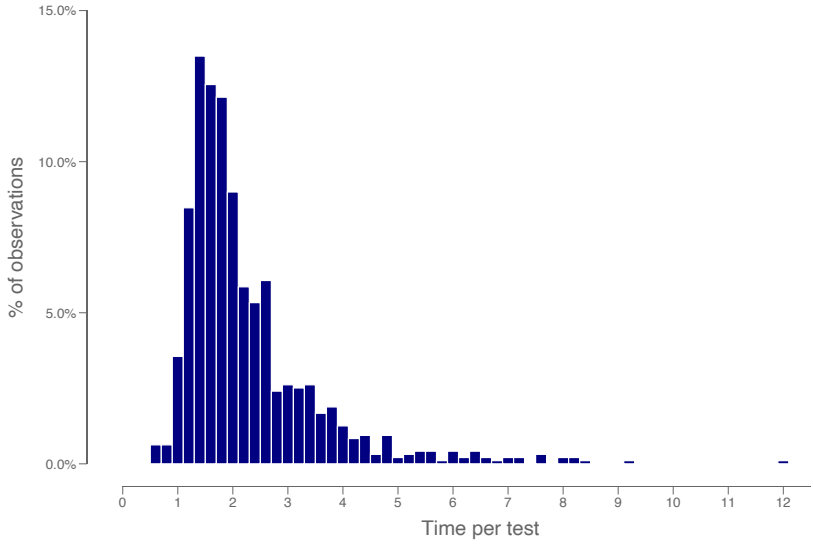


Figure 4.10: Distribution of the time per test measure. (Color version of this figure is available in the electronic version of this document.)

Chapter 5

On loan officers' decision strategies

And I thank you, I thank you for doing your duty
You keepers of truth, you guardians of beauty
— Leonard Cohen thanks those who hold him accountable

Summary:

An important part of many decisions is using imperfect cues to infer values on some criterion of interest, such as when a loan officer uses financial indicators and other information to gauge the creditworthiness of a company. Existing research demonstrates that people sometimes solve this task by relying on information-intensive strategies that attempt to integrate all available information and other times use information-frugal strategies that ignore some of the information. Despite their simplicity, information-frugal strategies can be ecologically rational and some research indicates that their use may be more common among expert decision makers. In this chapter we combine Bayesian latent-mixture modeling and analysis of eye-tracking data to explore the decision strategies of 42 professional loan officers. In contrast to comparable studies in other decision domains, our modeling results suggest that the majority of loan officers used a compensatory strategy. In particular, most subjects were classified as users of the information-intensive Weighted Additive strategy and only six subjects were classified as the users of the information-frugal Take The Best strategy. These results are corroborated and extended by the analysis of eye-tracking data. We conclude by identifying joint modeling of eye-tracking and decision data as a particularly promising avenue for future research.

Acknowledgments:

We thank the loan officers who participated in the study as well as our other contacts at the bank for their time and collaboration. We also thank Lucrezia Fattobene for the discussions that informed this research and Thorsten Pachur for providing us with the data from Pachur and Marinello (2013) and for his comments on a draft of this chapter.

5.1 Introduction

Decision making in the real world often involves the task of inferring a value on some *criterion* on the basis of imperfect *cues*. For example, loan officers attempt to gauge the creditworthiness of companies on the basis of financial cues, such as the ratio of cash to total assets, and non-financial cues, such as the quality and credibility of the management team. The strategies that people use in probabilistic inference are of great interest to fundamental disciplines, such as psychology and cognitive science, but also applied ones, such as management, law, and medicine (Artinger et al., 2015; Dhami and Ayton, 2001; Wegwarth et al., 2009). By documenting the strategies that decision makers use, researchers can systematically investigate the *ecological rationality* of these strategies (Todd and Brighton, 2016). That is, the environmental conditions under which the strategies perform better or worse can be investigated using real world data, computer simulations, and mathematical analysis. As an example, Luan et al. (2019) recently used both real and simulated data to investigate and contrast the accuracy and search effort of decision strategies in the context of personnel selection. In this way, research on decision strategies can make a significant contribution to practice.

Existing research makes it clear that people sometimes use information-intensive strategies that involve an exhaustive search for alternatives and cues. We use the term *alternatives* to describe decision options (e.g. approve the loan) and the term *cues* to describe the criteria for making a choice (e.g. the ratio of cash to total assets). Information-intensive strategies are sometimes described as *rational*, as they aim to make use of all available information and thus mimic or approximate normative theories of choice (Luan et al., 2019). They are also *compensatory*, as cues lower in importance can in principle compensate for cues higher in importance. A prototypical strategy of this class is the Weighted Additive strategy (Payne et al., 1993). This strategy determines the overall value of each alternative by multiplying every cue value with its respective weight and summing the resulting values. The alternative with the highest overall value is then selected. Information-intensive strategies like the Weighted Additive strategy can be contrasted with information-frugal strategies (often called *heuristics*) that simplify decision making by ignoring some of the alternatives or some of the cues. A prototypical strategy of this class is the Take The Best strategy (Gigerenzer and Goldstein, 1999). This strategy compares the alternatives by inspecting the cues sequentially in an order determined by their importance, starting with the most important cue. This strategy terminates search as soon as a discriminating cue is found and chooses the alternative with the better value on that cue. In this way, the Take The Best strategy typically substantially reduces the number of cues that need to be inspected compared to the Weighted Additive and other information-intensive strategies (Lee et al., 2017). Take The Best is a *noncompensatory* strategy, as cues lower in importance can never compensate for cues higher in importance.

The conditions under which people use information-intensive, compensatory strategies such as the Weighted Additive and information-frugal, noncompensatory ones such as the Take The Best have been a focus of intense interest in psychology. Most of these studies have focused on novice decision makers performing artificial tasks. But researchers have also begun to investigate how well these strategies capture the choices of experienced decision makers, such as magistrates making bail decisions, police officers and burglars assessing which property is more likely to be targeted for burglary, and airport customs officers judging which passengers

are more likely to be smuggling drugs (Dhami and Ayton, 2001; Garcia-Retamero and Dhami, 2009; Pachur and Marinello, 2013). We contribute to this stream of literature by investigating the use of these decision strategies among professional loan officers. Previous studies on loan officers' judgment and decision making have focused on loan officers' use of accounting information in predicting business failure (Libby, 1975; Casey, 1980; Abdel-Khalik and El-Sheshai, 1980), the effect of task characteristics on information acquisition and choice (Casey, 1980; Biggs et al., 1985), the role of different types of information in lending decisions (Beaulieu, 1994; Catasús and Gröjer, 2003), and the differences betweeen experienced and novice loan officers (Rodgers, 1999; Andersson, 2004).

What kind of decision strategies are likely to be the most prevalent among loan officers? We preface our attempt at answering this question by introducing two theoretical perspectives on strategy selection (Payne et al., 1993; Gigerenzer and Todd, 1999). The two perspectives share a foundational assumption that people possess a repertoire of decision strategies that they can draw upon. Both also recognize that information-frugal strategies form a part of the repertoire in addition to information-intensive ones. But the two perspectives differ in their explanation of *why* people sometimes rely on information-frugal strategies. One perspective emphasizes that decision making involves an *accuracy/effort trade-off* (Payne et al., 1993). Information-frugal strategies are thought to be less accurate than information-intensive ones while conferring the benefit of reduced effort (Shah and Oppenheimer, 2008). Their use can be rational if the accuracy gains of information-intensive strategies do not justify the increased costs in time, effort, and other resources (Gigerenzer and Gaissmaier, 2011). This assumption of the accuracy/effort trade-off is challenged by the research program on *fast-and-frugal heuristics* (Gigerenzer and Todd, 1999). These researchers point out that information-frugal strategies can in fact outperform information-intensive strategies even when it comes to accuracy, and emphasize the study of the *ecological rationality* of different strategies – that is, the study of environmental conditions that determine how well a given strategy can accomplish the decision maker's goals. Simulation studies using real world data show that information-frugal strategies such as the Take The Best strategy can outperform information-intensive strategies such as the Weighted Additive strategy (Czerlinski et al., 1999; Şimşek, 2013; Lee et al., 2017; Luan et al., 2019).

The two perspectives outlined above inspire two contrasting hypotheses concerning the decision strategies used by experienced loan officers. The perspective emphasizing the accuracy/effort trade-off of decision strategies suggests that experienced decision makers may be more likely to use information-intensive strategies, such as the Weighted Additive strategy. Experience in a domain is likely to affect both dimensions of the trade-off between accuracy and effort. Concerning accuracy, experience could contribute to adaptive strategy selection through both training and selection effects. Concerning effort, it is well-known that experience decreases effort by making initially deliberate operations automatic through repeated application. Experienced decision makers might thus be able to carry out more information-intensive strategies with less effort than novice decision makers. Jointly, these observations seem to point in the direction of information-intensive strategies being the norm among loan officers. On the other hand, complex real world decision environments such as commercial lending are characterized by limited outcome data, cues of questionable validity, and change. These are in fact the conditions under which information-frugal strategies seem to perform particularly

well (Gigerenzer, 2008). The ecological rationality perspective would therefore suggest that experienced decision makers in these environments would be particularly likely to be using information-frugal strategies suited to these environmental conditions. Indeed, some recent studies support this view. Both Garcia-Retamero and Dhami (2009) and Pachur and Marinello (2013) report that the information-frugal Take The Best was the most prevalent strategy among experienced decision makers in their studies. In fact, experienced decision makers were *more* likely to be using the Take The Best strategy than novice decision makers.

It is worth noting, however, that the decision domains examined by Garcia-Retamero and Dhami (2009) — burglary — and Pachur and Marinello (2013) — airport customs patrol — differ markedly from the domain of commercial lending. For example, airport patrol officers might make hundreds of quick decisions each day, and time pressure has been associated with the use of noncompensatory strategies such as Take The Best (Rieskamp and Hoffrage, 2008). In contrast, the pace of decision making is much slower in commercial lending. Another significant difference is that loan officers do not make decisions by themselves. Instead, the decision making process is intensely social and composed of multiple stages. The need to justify one's judgments and recommendations thus looms large in the case of commercial lending. The literature on accountability suggests that the need for justification facilitates exhaustive information gathering and compensatory decision making (Kahn and Baron, 1995; Lee et al., 1999). In line with this reasoning, Andersson (2004) found that professional loan officers engaged in extensive search for cues, and that the extent of search increased with experience.

In the current study, we use a modern Bayesian outcome-based classification analysis to investigate the decision strategies used by loan officers. We contrast the information-intensive and compensatory Weighted Additive strategy with the information-frugal and noncompensatory Take The Best strategy. In addition, and following a number of previous studies, we also add the Equal Weights strategy (Einhorn and Hogarth, 1975) to our set of candidate strategies. This strategy can be formulated as a special case of the Weighted Additive strategy with all cue weights set to the same value. It is a compensatory strategy that takes into account all alternatives and cues and is therefore sometimes classified along with the Weighted Additive strategy as a "rational" strategy compared to Take The Best (Lee et al., 2019). On the other hand, the Equal Weights strategy does ignore some information (specifically the cue weights) and may therefore also be classified as an information-frugal strategy (Gigerenzer and Gaissmaier, 2011). Against the backdrop of the contrasting theoretical perspectives introduced above, we do not make a sharp prediction about loan officers' decision strategy use, instead formulating our main aim as a research question.

Progress in decision research is facilitated if decision models (such as the strategies considered in this chapter) are specified as formal models (Marewski et al., 2010). Formalization contributes to both the descriptive aim (by enabling the use of formal statistical methods to infer the decision strategies used by subjects) and the prescriptive aim (by enabling the use of simulations to investigate the conditions under which the strategies perform better or worse) of decision research. The three strategies considered in this chapter are typically formalized by assuming that decision makers follow their predictions, save for some unsystematic error. Although the assumption of unsystematic error is an obvious first step, it does not appear psychologically tenable, particularly for information-intensive strategies such as the Weighted

Additive strategy. The development of more plausible error specifications is thus an important aim of research on decision strategies (Hilbig and Moshagen, 2014; Lee, 2016; Heck et al., 2017). In the current chapter, we contribute to this project by introducing a novel version of the Weighted Additive strategy that formalizes the insight that errors (deviations from what the strategy predicts) should become less likely as the overall value difference between the alternatives increases (see *Methods*).

Our main analysis of decision strategy use is based on the formal comparison of subjects' decisions and the implied predictions of decision strategies (Bröder and Schiffer, 2003; Lee, 2016). But although outcome-based classification approaches of this sort are the standard in the literature, they also have limitations (Lee et al., 2019). As a complement to these approaches, researchers are increasingly using so-called *process* methods (Schulte-Mecklenbeck et al., 2017). Classical process methods include the *closed information board* and its modern successor *mouse-tracking* as well as *think-aloud* protocols. In the literature on loan officers' decision making, these methods are used in the studies of Biggs et al. (1985) and Andersson (2004). In the current study, we complement our main analysis by collecting and analyzing *eye-tracking* data. Eye-tracking provides rich data and is relatively unobtrusive compared to mouse-tracking and think-aloud protocols (Horstmann et al., 2009). Because of these advantages and the increased availability of eye-tracking systems, the usage of eye-tracking to investigate higher cognitive processes and economic behavior is on the rise (Orquin and Mueller Loose, 2013; Lahey and Oxley, 2016; Lynch and Andiola, 2018; Meißner and Oll, 2019). In the current study, eye-tracking provides useful additional information because the different strategies make predictions concerning cognitive processing that can be tested using eye-tracking data. For example, the Take The Best strategy makes predictions about the order in which cues will be inspected, which is something that is easy to test using eye-tracking data. A less obvious example concerns the idea that the use of the Weighted Additive strategy is cognitively more demanding than the use of heuristic strategies. Previous eye-tracking research (Horstmann et al., 2009) suggests that this prediction can be tested by investigating the duration of *fixations* — periods of time during which the eye(s) remain relatively still (Holmqvist et al., 2011).

Finally, another aim of the current study was to investigate the effects of accountability on lending decisions. As noted above, the need to justify one's judgments and recommendations looms large in commercial lending, and Andersson (2004) highlighted accountability as a potential explanation of extensive information gathering he observed among senior loan officers in his study. Accountability has been the subject of many studies in psychology, particularly laboratory experiments featuring students as subjects (Lerner and Tetlock, 1999). Brown (1999) emphasizes the need for naturalistic investigations of accountability by arguing that the effects of accountability in business environments "diverge considerably from those found in the lab." We attempt to contribute to this project by including a naturalistic manipulation of accountability in our study and investigating its effect on a range of outcomes.

5.2 Methods

Subjects

Forty-two professional loan officers took part in the study. Thirty-five were male, average age was 36 years (SD = 10). Professional experience (number of years working as a loan officer) ranged from less than a year to more than 40 years, the average being 9 years (SD = 9). All subjects worked for the same commercial bank. Participation was voluntary and there was no financial compensation for taking part in the study. All subjects had normal or corrected-to-normal vision.

Materials

Decision task. Every trial of the task featured two alternatives (fictional companies) described by 12 cues. These cues were selected in consultation with a senior loan officer working at the same bank as the subjects and include *Cash to total assets*, *Current ratio*, *EBIT interest coverage*, *Funds from operations to total debt*, *EBITDA margin*, *Total debt to EBITDA*, *Corporate governance, key-man and reporting*, *Management track record & strategy*, *Cost structure & operational efficiency*, *Growth*, *Geographical diversification*, and *Market leadership*. The first six of these cues are financial indicators and continuous in nature. Subjects were informed that specific values on these cues were translated into five categories, namely *very conservative*, *conservative*, *moderate*, *aggressive*, and *very aggressive*. Furthermore, only two of these categories were to appear in the task; *conservative* (positive cue value) and *aggressive* (negative cue value). For the other six cues, subjects were told to assume that another loan officer had examined the company in question and reported a judgment of either *well above average*, *above average*, *average*, *below average*, or *well below average*. As with financial indicators, only two of these categories were to appear in the task; *above average* (positive cue value) and *below average* (negative cue value). Subjects' task was to indicate which of the two companies in a given trial was the better candidate for loan approval. To construct the set of 30 trials used in the study, we followed a similar procedure as Pachur and Marinello (2013). First, we generated a set of all possible trials, before excluding (i) trials where one alternative dominated the other, (ii) trials in which the two alternatives were equally attractive according to the Equal Weights strategy, (iii) trials in which the two alternatives were equally attractive according to the Weighted Additive strategy (with cue weights supplied by abovementioned senior loan officer), (iv) trials in which the Equal Weight and Take The Best strategies made the same prediction, and (v) trials in which the Weighted Additive and Take The Best strategies made the same prediction. From the remaining trials, we randomly sampled 10 trials in which the most important cue (according to the senior loan officer) discriminated between the alternatives, 10 trials in which the second most important cue discriminated between the alternatives (whereas the most important one did not), and 10 trials in which the third most important cue discriminated between the alternatives (whereas the first two did not). All subjects completed this same set of 30 trials but in random order (different for each subject). Figure 5.1 shows an exemplary trial. The placement of cues and alternatives stayed constant between trials. Because the placement on the screen can be consequential, we created four versions of the task which differed in

	Company A	Company B
Cash to total assets	conservative	conservative
Current ratio	aggressive	conservative
EBIT interest coverage	aggressive	conservative
Funds from operations to total debt	aggressive	conservative
EBITDA margin	conservative	conservative
Total debt to EBITDA	conservative	aggressive

	Company A	Company B
Corporate governance, key-man & reporting	below average	below average
Management track record & strategy	below average	below average
Cost structure & operational efficiency	above average	above average
Growth	below average	above average
Geographical diversification	above average	above average
Market leadership	below average	below average

Figure 5.1: Example trial of the main task.

the placement of the cues on the screen. For version 1, the exact cue placement can be seen in Figure 5.1. The screenshots for the other three versions are available in the online project repository (see *Additional information*).

Eye-tracking setup. We used the SMI RED-250 eye-tracker (sampling rate 250Hz) to collect the gaze data. Subjects were seated around 60 centimeters away from the monitor (resolution 1680 × 1050 pixels), and a chin rest was used to minimize head movement. The areas of interest were defined in correspondence with the rectangles seen in Figure 5.1.

Accountability manipulation. Subjects in the *high* accountability condition were instructed to make decisions as if they were part of their job and as if they would have to explain and defend them in front of the credit committee. They were also informed that the researcher would ask them to explain the reasoning behind their decisions. Subjects in the *low* accountability condition were instructed to follow their own, personal jugment and told that their decisions were not going to be evaluated by others.

Postexperimental questionnaire. The postexperimental questionnaire contained three parts. The first part asked the subjects to rank and rate the cues in order of their importance. For the ranks, no ties were allowed. For the ratings, the scale went from 1 (completely unimportant) to 100 (extremely important). In some cases, subjects provided ranks and ratings that were not fully consistent. These subjects were asked to make the necessary changes to achieve consistency. In two cases, the researcher failed to spot the inconsistency, and in one case the subject refused to change his or her initial (and inconsistent) responses. The second part of the postexperimental questionnaire asked the subjects to rate (on a 1-7 Likert scale) how they perceived each cue. One question asked perceived objectivity, the other about perceived precision. The third part of the postexperimental questionnaire contained a few demographic

questions.

Procedure

After being welcomed and signing the informed consent statement, subjects read the instructions on paper and were given the opportunity to ask questions. Next, they were seated in front of the eye-tracking computer where they completed the eye-tracking calibration procedure and one practice trial. Right before the main task, subjects read the text of the accountability manipulation. Subjects then completed the first half of the main task. Each trial began with a blank screen (500 ms), followed by a fixation sign in the center of the screen (2000 ms), followed by the decision information. Subjects were instructed to press any key after they had made their decision, which they then indicated with a mouse-click. Following the first 15 decisions, there was a short break. Afterwards, the calibration procedure was repeated and subjects then completed the second half of the task. Finally, subjects completed the postexperimental questionnaire on paper. Each session lasted about 45 minutes and featured a single subject.

Model for inferring decision strategies

An important goal of the current study was to infer the decision strategies underlying the decisions of loan officers. For this purpose, we used a Bayesian latent-mixture modeling approach (Lee and Wagenmakers, 2014; Lee, 2016). In the current section, we motivate and describe the model that we used. Our starting assumption was that the observed decisions are produced not by a single process but by several processes, which in our case correspond to three distinct decision strategies. Following Pachur and Marinello (2013) and others, we assumed that the strategies potentially used by our subjects were the Weighted Additive strategy, the Take The Best strategy, and the Equal Weights strategy. Whether a given strategy was used is not something that is directly observable. Instead, it can be inferred on the basis of the fit between the predictions of the strategy and the observed decisions. For the Equal Weights strategy, predictions could be made *a priori* and were the same for all subjects. In contrast, the predictions of the Weighted Additive strategy depend on cue weights, which may differ among subjects. We therefore used the data from the cue importance rating task to obtain cue weights, which were used to generate individualized predictions for the Weighted Additive strategy. Similarly, we used the data from the *ranking* task to generate the predictions for the Take The Best strategy.

The three strategies included in the model specify which alternative should be chosen on any given trial [1]. But actual human decisions never conform perfectly to known decision strategies. To increase the strategies' psychological plausibility, sharp predictions can be transformed to probabilities. The simplest approach is to assume that predictions are perturbed by random error (Bröder and Schiffer, 2003). We use this approach for Equal Weights and Take The Best strategies. The assumption of random error seems relatively more problematic for the

[1] It is also possible that a strategy does not predict a specific choice in a given trial. In our study, this can only happen for the Weighted Additive strategy and if the subject's importance ratings are such that on a particular trial the value of the alternatives is equal.

more complex Weighted Additive strategy, which requires the integration of all cue values and weights. Consider that the cue weights were provided retrospectively by our subjects. To the extent that these self-reported weights correlate imperfectly with "true" weights, apparent errors may arise. But, crucially, these errors will not be random. Instead, they will be more likely to arise when the difference in the weighted sums of the two alternatives is small. Arguments of this sort motivate the development of new versions of the Weighted Additive strategy. One flexible approach is to classify the trials according to the size of the difference in value between the two alternatives. Separate error parameters are then estimated for each such class, with the restriction that larger differences in value correspond to lower (or equal) probabilities of error (Hilbig and Moshagen, 2014; Lee, 2016; Heck et al., 2017). In our study, this approach would have been too flexible as there were many classes of trials that featured only a single trial. For this reason, we developed a novel probabilistic version of the Weighted Additive strategy. The basic idea is to express the absolute value difference between the two alternatives as a proportion of the maximum possible absolute value difference (obtained by summing the weights) and map the resulting proportion to the probability of making an error.

The model is formalized in Figure 5.2. We use the graphical notation of Lee and Wagenmakers (2014) in which nodes are used to represent variables and arrows are used to indicate dependencies between variables. The shape, color, and border of the variables convey information. The shape indicates whether the variable is *discrete* (squares) or *continuous* (circles), the color indicates whether the variable is *observed* (grey) or *latent* (white), and the border indicates whether the variable is *deterministic* (double-bordered) or *stochastic* (single-bordered). Finally, the two plates (one for subjects, one for trials) are used to indicate replication.

To make sense of the model, we start with the central node y_{ij}, which captures the decision made by subject *i* on trial *j*. It takes a value of 1 if company A was chosen and a value of 0 if company B was chosen. The probability that company A will be chosen depends, first, on ζ_i. This parameter indexes the strategy used by subject *i* and is given a uniform prior, meaning that, before the data are taken into account, the model considers the three strategies equally likely to be used. Second, the probability that company A will be chosen depends on the probability of "error", captured by the node ε_{ij}. The value of ε_{ij} in turn depends on the strategy used by the subject. For users of Equal Weights and Take The Best strategies, a single parameter is estimated. For users of the Weighted Additive strategy, however, the ε_{ij} varies by trial and depends on the value difference between the two alternatives (Diff_{ij}) as well as on the α_i parameter, which controls the shape of the relationship between Diff_{ij} and ε_{ij}. Examples of how this relationship depends on the value of α are visualized as yellow dashed lines in Figure 5.5.

We implemented the model using the JAGS software (Plummer, 2003) and collected 20000 samples from each of three chains, after a burn-in period of 5000 samples and after thinning so that only each fifth sample was kept. We checked that the model converged by visually inspecting the chains and using the $\hat{R}$ statistic (Brooks and Gelman, 1998).

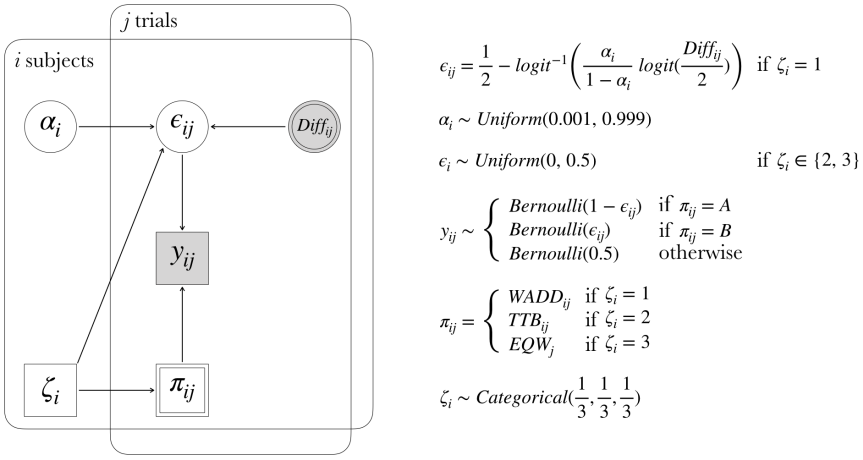


Figure 5.2: The decision-making model. A graphical representation of the Bayesian latent-mixture model used to infer the subjects' decision strategies.

Preprocessing of eye-tracking data

The raw eye-tracking data were preprocessed using the default velocity-based algorithm in SMI BeGaze (v3.5.101). The eye-tracking analyses were conducted on the fixations (minimum duration: 50 miliseconds) identified by the algorithm. The term *fixations* refers to periods of time during which the eye(s) remain relatively still (Holmqvist et al., 2011). The two attributes of fixations that are of interest are (i) their duration (informing us about the kind of cognitive processing the subject is engaged in) and (ii) the associated area on the computer screen (informing us about the information the subject is attending to).

Additional information

Additional information, including study materials and data (with the exception of demographic data), is available in the project's online repository (`https://osf.io/dc3nj/`).

5.3 Results

Preliminary analyses

We standardized the ratings of cue importance to a mean of zero and a standard deviation of one (computed separately for each subject) to increase comparability between subjects. The standardized ratings are displayed in Figure 5.3. Blue circles represent individual ratings and

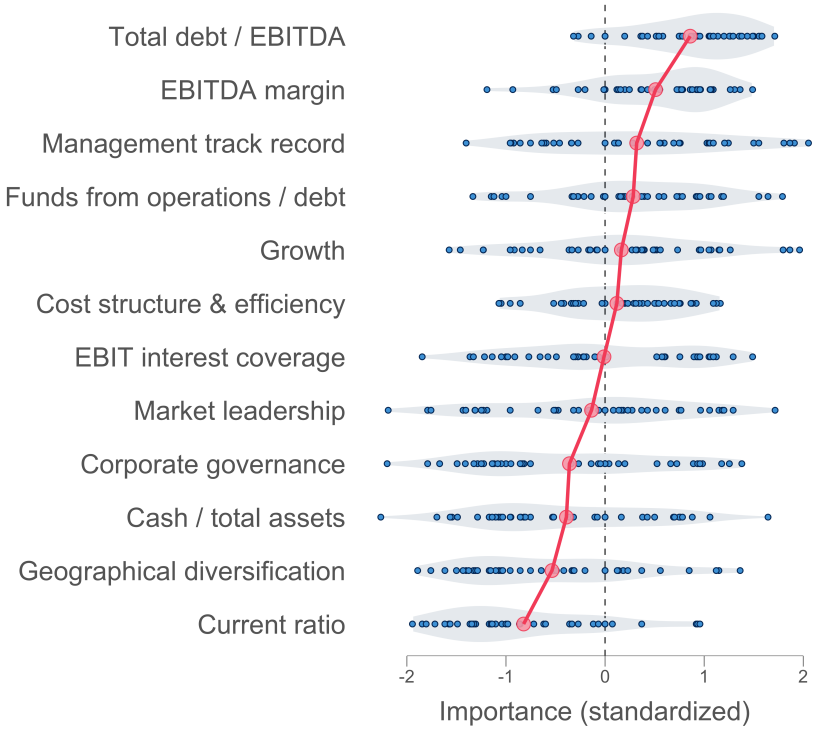


Figure 5.3: Standardized importance ratings for the 12 cues used in the study. The raw responses were standardized by subtracting the subject-specific mean and dividing the resulting score by the subject-specific standard deviation. Blue circles represent individual ratings, shaded areas track the densities of the distributions of ratings, and the connected red circles show the averages. (Color version of this figure is available in the electronic version of this document.)

the filled areas represent the estimated distributions of ratings. The averages (represented by the larger, red circles) correlate well with the ratings provided by a senior loan officer before the study (Pearson's $r = +0.78$). But we also observe considerable variability in the assessed relative importance of cues. For example, no cue was consistently rated by everyone as either above or below average in importance. The variation in these ratings is contrasted with substantial agreement in the actual choices made during the main task. Across the 30 trials, subjects on average picked the same decision alternative in 79% of the cases (range 60% - 98%), which is slightly larger than the 72% (range 52% - 94%) observed by Pachur and Marinello (2013) for airport customs officers in a comparable study.

Decision strategy classification

We fit the model described in the section *Model for inferring decision strategies*[2]. This model classifies subjects as users of the Weighted Additive strategy, the Equal Weights strategy, or the Take The Best strategy on the basis of the fit between subjects' decisions and the predictions of these strategies. The results are summarized in Figure 5.4. Each row corresponds to one subject and its color represents the inferred strategy. The majority of subjects were classified as users of a compensatory strategy, with 62% classified as Weighted Additive strategy users and 21% as Equal Weights strategy users. Only 14% of subjects were classified as users of the noncompensatory Take The Best strategy. For the remaining subject, the strategies Take The Best and Equal Weights had the exact same fit. These results contrast with those of Garcia-Retamero and Dhami (2009) and Pachur and Marinello (2013), where the Take The Best strategy provided the best fit for the majority of expert subjects.

Instead of taking the classifications at face value, however, it is important to consider the associated uncertainty. In Bayesian statistics, the Bayes Factor (Kass and Raftery, 1995) is often used as a measure of evidence. It is a ratio of how likely the given data are under one hypothesis compared to another. In this case, the "hypotheses" are the three possible classifications for each subject. We therefore computed two Bayes Factors for each subject, comparing the fit of the inferred strategy with each of the remaining two strategies. The results are summarized on the right side in Figure 5.4 and were also used to order the subjects along the y-axis. For about one half of the sample, we see moderate or strong evidence for the inferred decision strategy. For the other half, the Bayes Factor comparing the two best-fitting strategies is less than three, which is typically described as weak or anecdotal evidence (e.g. Raftery, 1995). For the vast majority of these subjects, however, there is clear evidence at least with respect to whether they used a compensatory strategy (Weighted Additive, Equal Weights) or a noncompensatory one (Take The Best). We conclude that the decisions of the majority of loan officers are more in line with the use of compensatory strategies than with the use of the noncompensatory Take The Best strategy.

In addition to the uncertainty associated with a given classification, it is important to consider the raw fit between subjects' decisions and the predictions of the inferred strategy. Figure 5.4 supplies this information as well. It shows, for each subject, how often they decided in accordance with the inferred strategy (highly saturated cells) and how often they did not (cells with low saturation). The mean accordance rate is 84% but there is significant variation among subjects. One subject that clearly stands out was classified as a user of the Weighted Additive strategy but more often than not made the opposite decision to the one prescribed by the strategy.

[2] We also fit two additional models. The first additional model was a version of the main model in which the Weighted Additive strategy had the same random probability of error as Take The Best and Equal Weights strategies. Overall, the results of this alternative model are similar to the results for the main model. One important difference is that with the alternative model, the strategies Weighted Additive and Equal Weights had the same fit for a subset of 11 subjects. The second additional model included just the two versions of the Weighted Additive strategy. Among the users of Weighted Additive strategy identified by either the main model or the first additional model described in this footnote, this alternative model assigned about 20% of the subjects to the version of the strategy with a random probability of error. The corresponding Bayes factors were mostly in the anecdotal range, however. In summary, the new version of the Weighted Additive strategy introduced in this chapter produced a slightly better fit for the majority of identified Weighted Additive strategy users than the traditional version with a random probability of error.

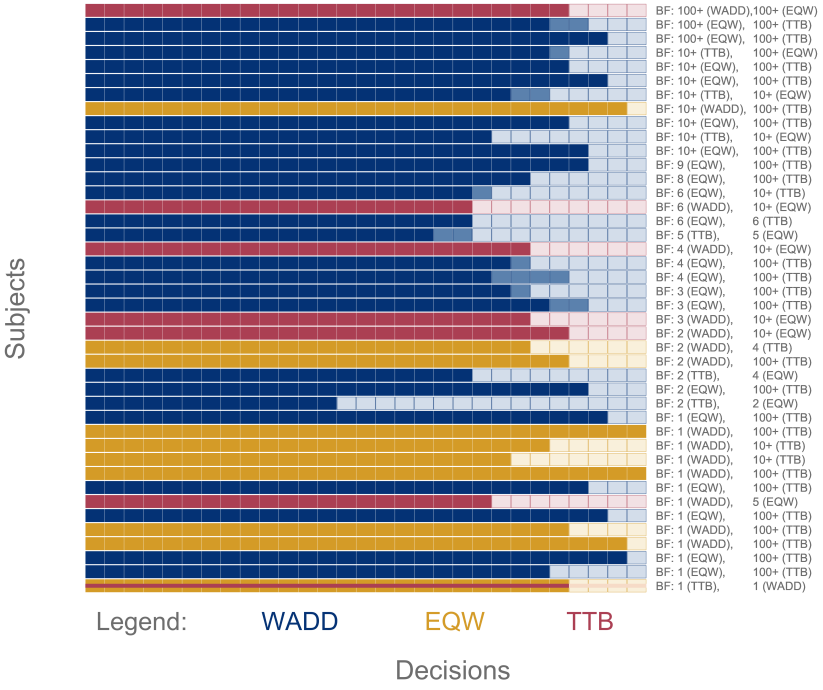


Figure 5.4: Strategy classifications and associated accordance rates and Bayes Factors. Each subject is represented by a single row, colored according to the strategy to which the subject was classified. The thirty columns represent thirty trials (order is not significant). High saturation of a cell signifies that the subject decided in accordance with their inferred strategy, low saturation signifies that they did not. In some trials, the two alternatives were deemed equally attractive by the Weighted Additive strategy; tese trials are displayed in an intermediate saturation. The labels on the right communicate two (rounded) Bayes Factors. The first (second) Bayes Factor compares the inferred strategy with the strategy that has the second (third) best fit. For the subject represented by the bottom row, the Equal Weights and the Take The Best strategies had the exact same fit. (Color version of this figure is available in the electronic version of this document.)

Figure 5.5 provides further insight for the subset of subjects who were classified as users of the Weighted Additive strategy. What is shown is the implied relationship between the absolute value difference of the two alternatives and the probability of making an "error" (i.e. not choosing the alternative favored by the Weighted Additive strategy). The subject that stood out in Figure 5.4 stands out here as well (the concave curve on the right). For this subject, none of the three strategies included in the model provide an acceptable account of his or her decisions. Alternatively, the cue weights supplied by the subjects may be particularly inaccurate.

Eye-tracking analyses

Unique inspections. We investigated to what extent the 24 decision-relevant pieces of information (12 cues × 2 alternatives) were inspected at least once in each trial. According to the Weighted Additive strategy and the Equal Weights strategy, (close to) all information should be inspected in each trial. In contrast, the information-frugal Take The Best strategy would inspect a much more limited amount of information. For example, for the cue ordering supplied by the senior loan officer before the study, between one and three cues (3 - 6 pieces of information) would be inspected per trial. The results show that the mean number of fully inspected *cues* was 10.6 for the average subject (range: 7.3 - 11.9). These results are consistent with the modeling results reported above that suggested that the majority of subjects used a compensatory strategy. Among the subjects classified as users of the Take The Best strategy, the mean number of fully inspected cues was lower (9.3, range: 7.6 - 11.7) but still relatively high compared to what we would expect if subjects adhered strictly to Take The Best's stopping rule.

Dwell time and fixation count. Fixation count refers to the total number of fixations to an area of interest, and dwell time refers to the summed duration of these fixations. Because of its stopping rule, the Take The Best strategy predicts a positive correlation between these measures and the importance of a cue. For example, the most important cue would be inspected every time whereas the least important cue would be inspected only if the preceding 11 cues all failed to discriminate between the two alternatives. In contrast, there is no reason to expect a positive correlation for users of the Equal Weights strategy. Finally, for users of the Weighted Additive strategy the prediction is less clear and depends on assumptions about the cognitive implementation of the strategy. To test these predictions, we computed subject-level correlations between dwell time and self-reported cue importance [3]. The results are shown in Figure 5.6 and reveal a positive correlation for 87% of the subjects[4]. As expected, the mean correlation is higher for Take The Best users (+0.55) than for Equal Weights users (+0.21), with Weighted Additive strategy users in between (+0.37).

Information search. Previous analyses revealed the extent to which the decision-relevant information was attended to and how the amount of attention to a cue varied as a function of its reported importance. Another important aspect of information processing revealed by eye-tracking is the type of visual search subjects engaged in. We investigated two dimensions

[3] The results are similar if fixation count is used instead of dwell time.

[4] The results for three subjects are not shown because the lack of variation in their importance ratings precluded the computation of the correlation coefficient.

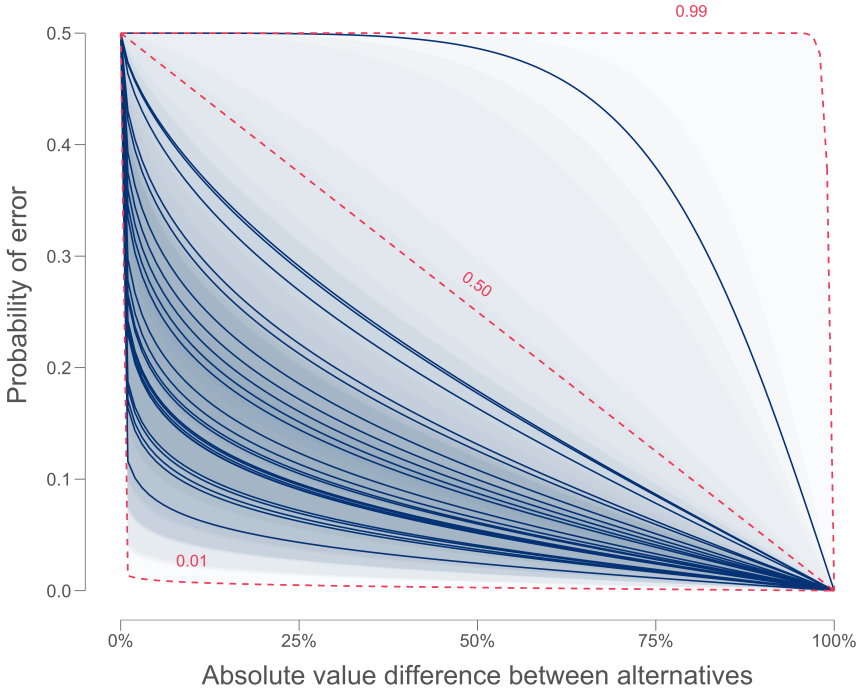


Figure 5.5: Model-implied relationship between the absolute value difference between the two alternatives and the probability of error (not choosing the alternative that the strategy predicts) for subjects classified as users of the Weighted Additive strategy. Each subject is represented by one purple line (constructed using the posterior means of the alpha parameter) and one confidence band (constructed using the 95% uncertainty interval). Confidence bands are drawn with high opacity to prevent overplotting. The three red, dashed lines illustrate how the value of the alpha parameter translates to the shape of the curve. (Color version of this figure is available in the electronic version of this document.)

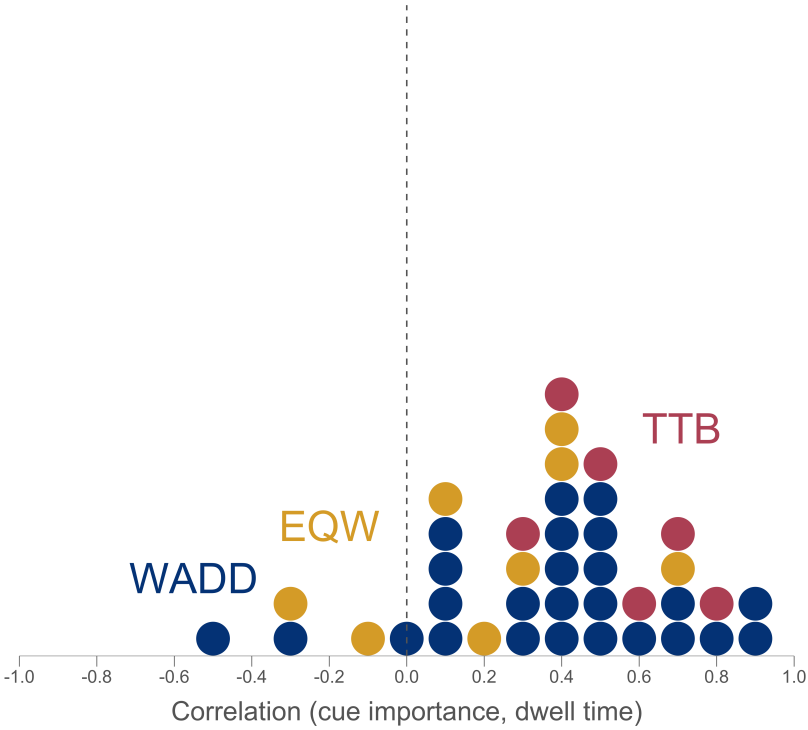


Figure 5.6: Correlation between cue importance and dwell time. Each circle represent one subject and its color conveys the decision strategy that the subject was classified as using (WADD = Weighted Additive strategy, EQW = Equal Weights strategy, TTB = Take The Best strategy). The x-axis represents Pearson's correlation between the importance assigned to a given cue and the dwell time for that cue. (Color version of this figure is available in the electronic version of this document.)

of search behavior. We begin by explaining these dimensions and the necessary calculations before presenting the results jointly in Figure 5.7.

The first dimension concerns the distinction between *cue-based* and *alternative-based* processing. The former corresponds to comparing the alternatives on a given cue before proceeding to the next cue. The latter corresponds to investigating a single alternative on several cues before moving to the other alternative. The mapping between a strategy and type of processing thus defined is clear for Take The Best, which prescribes cue-based processing. To develop predictions for the other two strategies, we first assume that subjects choose the type of processing that is cognitively least costly. Consider the Weighted Additive strategy. Even though both cue-based and alternative-based processing can support it, the former is clearly cognitively easier. That is because the subject can simply avoid processing the cues which do not discriminate between the two alternatives. In contrast, these cues would need to be processed under alternative-based processing. I therefore predict cue-based processing for users of the Weighted Additive strategy. For the Equal Weights strategy, in contrast, we assume that alternative-based processing would be more likely because the Equal Weights strategy translates into a simple count of positively-valued cues. The reader is invited to test the reasonableness of these assumptions using Figure 5.1.

To investigate whether subjects engaged in cue-based or alternative-based processing, we begin by defininining *transition* as the pattern of information search contained within a pair of successive fixations. If we first eliminate fixations that do not fall on areas of interest containing cue values, four types of transitions are possible. These are defined by *(i)* whether the second fixation landed on the same *cue* (e.g. *Growth*) as the first one and *(ii)* whether the second fixation landed on the same *alternative* (e.g. *Company A*) as the first one. Ignoring successive fixations to the same area of interest leaves us with *within-cue* (same cue, different alternative), *within-alternative* (different cue, same alternative), and *shift* transitions (different cue, different alternative). Cue-based processing consists of patterned search behavior where each pattern begins with one or more within-cue transitions followed by either a within-alternative or shift transition (as the user moves to the next cue). It therefore predicts that the proportion of within-cue transitions will be about one half. In contrast, alternative-based processing consists mainly of within-alternative transitions and therefore predicts a low proportion of within-cue transitions.

The second dimension concerns the order in which cues were inspected. We identified two candidate types of search. *Importance-based* search examines cues in order of their importance. This is the type of search prescribed by the Take The Best strategy. In contrast, *position-based* search is based on the spatial arrangement of cues on the screen. In our case, position-based search would start in the top left corner, proceed downwards until the first "column" of cues is examined (see Figure 5.1), and then move to the top right corner and again proceed downwards until the second column of cues is examined. This prediction is based jointly on traditional reading direction for Western subjects and the arrangement of cues on the screen. Position-based search is clearly most likely for users of the Equal Weights Strategy, as it minimizes cognitive costs. For the Weighted Additive strategy, an argument could be made for both importance-based and position-based search; a sharp prediction is therefore not warranted. To analyze subjects' search orders, we classified each trial as position-based or importance-based. To do this, we considered only fixations targeting areas of interests

containing cue values. Next, we disregarded the information about the *alternative* which was attended to and removed repetitions. We recoded the resulting fixation sequences as strings. This enabled the use of the *Levenshtein distance* to check if the actual sequence was closer to the sequence implied by position-based search or the sequence implied by the importance-based search[5]. The Levenshtein distance is the minimal number of character insertions, deletions, and substitutions needed to transform one string into another. If the Levenshtein distance for the position-based order was smaller than for the importance-based order, a trial was classified as position-based — and vice versa.

Figure 5.7 displays the results. The mean proportion of within-cue transitions is shown on the x-axis and ranges between 19% and 54%. The data suggest that the majority of subjects engaged in a mixture of cue-based and alternative-based processing. Concerning cue search, we find that position-based order predominates. Forty-five percent of the subjects seemed to rely exclusively on position-based order and 93% relied on it more often than not. Only three subjects used importance-based search in the majority of trials. Finally, we do not observe striking differences between the users of different decision strategies. Even among Take The Best users, position-based search predominates, although the mean proportion of trials with position-based search is higher for users of Weighted Additive (89%) and Equal Weights (88%) strategies than for Take The Best users (77%).

Fixation duration. We computed the mean fixation duration for each subject and display the results in Figure 5.8. To the extent that the more complex Weighted Additive strategy requires more deliberate decision making than the heuristic Equal Weights and Take The Best strategies, we would expect to see longer fixations among its users. The data displayed in Figure 5.8 do not support this prediction, however. It may be that the computations enabling the Weighted Additive strategy are in fact performed in a more automatic manner. Alternatively, genuine differences between the users of different strategies could have been masked by our use of averaging. Users of the Weighted Additive strategy could have a higher number of longer fixations (corresponding to more deliberate processing) but also a larger number of shorter fixations. To address this possibility, we classified fixations as short (< 150 ms), medium (150 - 500 ms), or long (500+ ms) (Horstmann et al., 2009), and display their average number for each subject in Figure 5.9. The results are highly similar across the three strategies. Fixations of medium durations predominate, followed by short fixations; long fixations are rare. Overall, then, we found that duration of fixations does not discriminate between users of different strategies.

Accountability

The distribution of inferred decision strategies was similar in the two accountability conditions. In the low accountability condition, 62% of the subjects were classified as users of the Weighted Additive strategy, 29% as users of the Equal Weights strategy, and 10% as users of the Take The Best strategy. In the high accountability condition, 62% of the subjects were classified as users of the Weighted Additive strategy, 19% as users of the Equal Weights strategy, and 19% as users of the Take The Best strategy. To test for other differences between the two

[5] To obtain unique strings for importance-based search, we used the cue *ranking* data instead of the cue *rating* data.

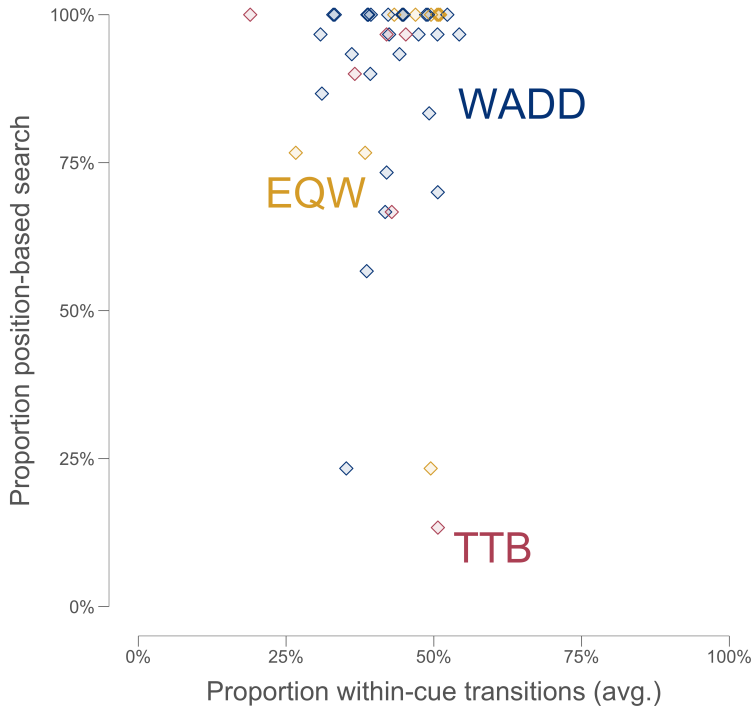


Figure 5.7: Two dimensions of search behavior revealed by eye-tracking. Each diamond represents one subject, and the color represents the inferred strategy (WADD = Weighted Additive strategy, EQW = Equal Weights strategy, TTB = Take The Best strategy). The *x*-axis shows the average proportion of within-cue transitions in a trial. The *y*-axis shows the proportion of trials in which the movement between cues was classified as position-based (as opposed to importance-based). (Color version of this figure is available in the electronic version of this document.)

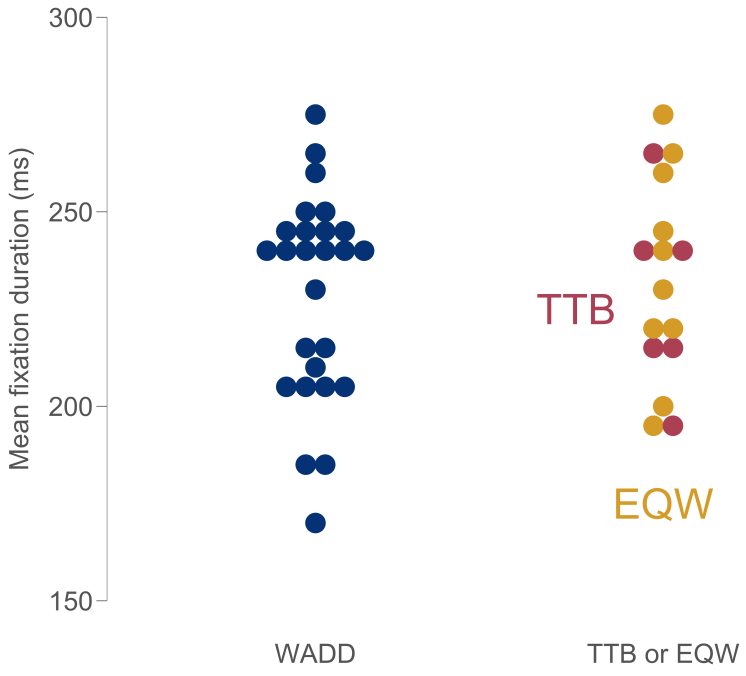


Figure 5.8: Mean fixation duration (in miliseconds) for users of different strategies. Each circle corresponds to one subject, with observations binned together (bin width = 5 miliseconds). WADD = Weighted Additive strategy, TTB = Take The Best strategy, EQW = Equal Weights Strategy. (Color version of this figure is available in the electronic version of this document.)

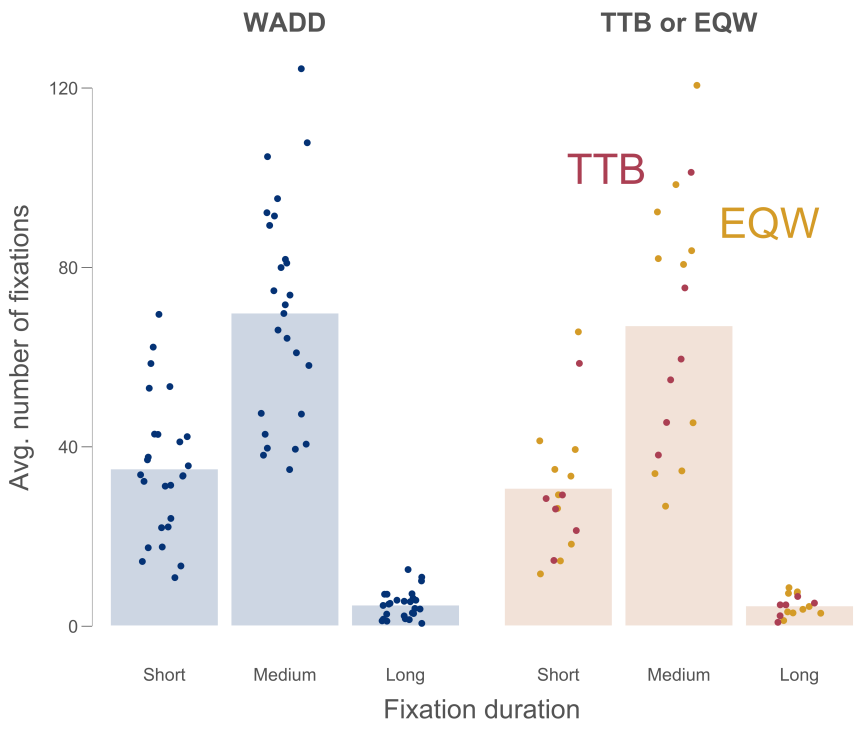


Figure 5.9: Average number of short, medium, and long fixations for each subject. The bars represent mean values, the dots represent individual subjects. The color of each dot conveys the strategy that was inferred for that subject. WADD = Weighted Additive strategy, TTB = Take The Best strategy, EQW = Equal Weights strategy. (Color version of this figure is available in the electronic version of this document.)

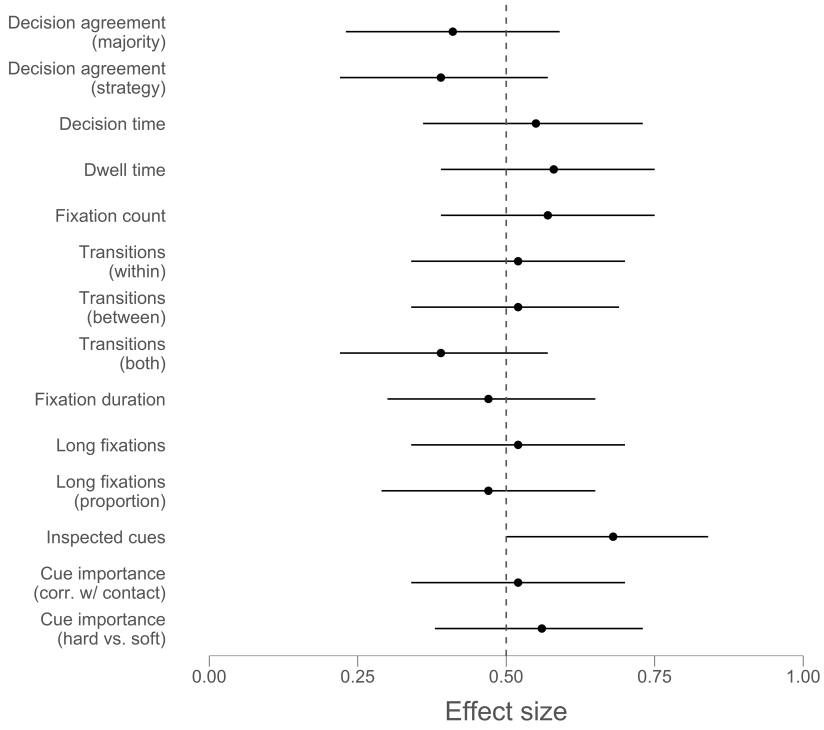


Figure 5.10: Comparison of the two accountability conditions for all major measures. The effect size on the *x*-axis refers to the probability (obtained using simulation) that a randomly selected subject from the high accountability group will have a larger score on a given measure than a randomly selected subject from the low accountability group. The dots correspond to the estimates and the lines represent 95% uncertainty intervals.

accountability conditions, we identified 14 measures of interest. These include measures based on eye-tracking data, measures based on choice and response time data, and measures based on cue importance data. For each measure, we computed the probability that a randomly selected subject from the high accountability condition will have a higher score than a randomly selected subject from the low accountability condition (see McGraw and Wong, 1992). We then used a bootstrapping procedure with 5000 iterations to construct 95% uncertainty intervals around these estimates. The results are plotted in Figure 5.10. Although the data provide some indication that the extent of information search (unique inspections) increased under high accountability, the differences between the two accountability conditions are generally small relative to the associated uncertainty, precluding any clear conclusion regarding the effect of the accountability manipulation.

5.4 Discussion

The main aim of the current study was to investigate the decision strategies used by professional loan officers in a probabilistic inference task. In each trial of the task, subjects inspected the profiles of two fictional companies and indicated which of the two companies seemed a better candidate for loan approval. We used these decisions as an input to a Bayesian latent-mixture model which classified each subject as a user of one of three decision strategies that have been previously found to provide a good account of decision making in comparable tasks. The results indicate that the majority of subjects used a compensatory strategy. In particular, the Weighted Additive strategy was inferred for more than half of the sample[6]. This finding resonates with many laboratory studies on probabilistic inference that, like our study, did not feature explicit information acquisition costs, time pressure, or immediate accuracy feedback (e.g. Heck et al., 2017). But it contrasts with the results of similar studies on expert decision making in other domains. For example, Garcia-Retamero and Dhami (2009) found that the majority of decisions made by police officers and experienced burglars were best predicted by the Take The Best strategy. Similarly, the majority of airport customs officers in the study by Pachur and Marinello (2013) were classified as Take The Best users. In comparison with these domains, the domain of commercial loan lending is characterized by the large amount and complexity of information, slower pace of decision making, lack of fast and unambiguous feedback, and the need to justify one's judgment to others. These factors seem likely to incentivize the usage of information-intensive, compensatory decision strategies (Andersson, 2004). Even if information-frugal strategies would simplify decision making at no cost to accuracy (Czerlinski et al., 1999; Şimşek, 2013; Lee et al., 2017; Luan et al., 2019), some of these same factors could prevent the discovery of this possibility.

In the current study we recorded and analyzed eye-tracking data in order to gain further insight into the similarities and differences among the users of different decision strategies. The results based on eye-tracking data resonate with the results obtained using decision data. For example, subjects engaged in an exhaustive search, typically inspecting all or almost all of the information presented on the screen. Similarly, Andersson (2004) reported extensive information gathering in a study involving Swedish loan officers. In the current study, the average number of fully inspected cues was slightly lower for users of the Take The Best strategy but still relatively high. This is in line with other research showing that although the Take The Best strategy sometimes provides a good account of people's *choices*, its *information search* predictions fare less well descriptively (Dummel et al., 2016). Concerning the *type* of search subjects engaged in, eye-tracking data shows that most subjects seem to have employed a mixture of cue-based and alternative-based processing, and search order was most often determined by the arrangement of the cues on the screen. The prediction of the Take The Best strategy that subjects will tend to inspect cues in order of their importance held true for three subjects, and only one of them was classified as a Take The Best user. Where observed, the differences between users of different strategies were in line with expectations but relatively minor. For example, mean correlation between cue importance and the associated dwell time was highest among Take The Best users (as expected given its stopping rule) but positive across

[6] It should be noted, however, that in a substantial minority of cases the Weighted Additive strategy and the Equal Weights strategy had a very similar fit.

all three strategies. It is important to note, however, that the number of subjects classified as Take The Best users was low. Because we would expect the strongest differences to be between users of Take The Best and users of compensatory strategies, our observations need to be replicated in a study that would have a more even split between these strategies.

The final aim of the study was to investigate the effects of a naturalistic accountability manipulation on predecisional and choice behavior. As previously discussed, accountability looms large in Lending decisions, and Andersson (2004) identified accountability as a possible explanation of extensive information gathering he observed among senior loan officers in his study. We assigned subjects to either a low accountability or a high accountability condition and compared the two groups on a variety of outcomes. The results, however, were not precise enough to support any clear conclusion. We did observe some evidence that the average number of fully inspected cues increased under high accountability, a finding which would be in line with the speculation of Andersson (2004) and the general accountability literature (Lerner and Tetlock, 1999; Lee et al., 1999). Given the associated uncertainty, however, we cannot read too much into this. Our lack of clear results concerning accountability can be likely traced to the low sample size and the hypothetical nature of our manipulation. It is also possible that the effects of accountability are more internalized and therefore less responsive to situational manipulations (Russo et al., 2000), or that the effects vary significantly due to the different social constituencies of different loan officers (Brown, 1999).

Before considering promising avenues for future research, we note two additional limitations of the current study. First, there is a mismatch between the type of task we have used (decision between two alternatives) and the type of task primarily faced by loan officers in their work (evaluation of a single alternative). Previous research shows that these two modes of decision making are not always exchangeable (Hsee, 1996). Second, the information conveyed by the cues in our task was simplified significantly to accommodate methodological and practical constraints. For example, cue values were categorized instead of being presented as exact values (in the case of financial cues) or in some even richer form (in the case of non-financial cues).

The current study highlights several promising avenues for future research. The version of the Weighted Additive strategy that we have developed needs to be investigated further. In the current study, this version of the strategy provided a better fit for the majority of relevant subjects than the traditional version with a random probability of error. For most of these subjects, however, the evidence in favor of one or the other strategy was in anecdotal range. It would be instructive to compare the two versions of the strategy in a study in which cue weights were provided *a priori* and therefore known with greater precision than in the current study, where the weights were estimated by subjects themselves. Another important next step would be to also develop versions of the Equal Weights and Take The Best strategies that incorporate more plausible theories of error (Heck et al., 2017). Finally, in our view a very promising direction for future research involves the integration of eye-tracking and decision data. In the current chapter, we have classified subjects as users of different decision strategies on the basis of decision data, and followed up by using eye-tracking data to highlight the similarities and differences between users of different strategies. A natural next step would be to formally include in the model visual behavior as captured by eye-tracking, thus increasing the scope of the model (Rieskamp and Otto, 2006; Glöckner, 2009; Lee et al., 2019). Such a

model could also include the possibility of strategy switches during the task (Lee, 2019). In a recent reanalysis of Walsh and Gluck (2016), Lee et al. (2019) found strong evidence for strategy switching in a large majority of subjects completing a probabilistic inference task. The development of these more complete models will bring us closer to the challenging but important task of understanding the decision making of loan officers and other significant economic agents.

5.5 Appendix 1

Table 5.0 provides the definitions of variables used in the figures. For Figure 5.10 (rows 6-16 in the table), the effect size is the probability that a randomly selected subject from the high accountability condition will have a higher score than a randomly selected subject from the low accountability condition. What is defined in the table are the underlying measures on which this effect size is based.

Table 5.0: Variable definitions.

Label	Definition [relevant figure]
Importance (standardized)	Cue importance ratings scaled, separately for each subject, to have a mean of 0 and a standard deviation of 1. [Fig. 5.3]
Correlation (cue importance, dwell time)	Correlation between the subject's cue importance ratings and dwell time for that cue. Dwell time (computed from fixations) measures how long, on average, the subject attended to that cue. [Fig. 5.6]
Proportion within-cue transitions	Average proportion of within-cue transitions. Transition is a sequence of two successive fixations. "Within-cue" means that the two fixations fell on the same cue but on different alternatives. [Fig. 5.7]
Proportion position-based search	Proportion of trials classified as involving position-based search rather than importance-based search. Position-based search roughly means moving from the current cue to the next closest cue. Importance-based search means moving to the next most important cue. [Fig. 5.7]
Mean fixation duration	Average duration of fixations in miliseconds. [Fig. 5.8]
Decision agreement (majority)	Proportion of trials where the decision matched the majority decision. [Fig. 5.10]
Decision agreement (strategy)	Proportion of trials where the decision matched the prediction of the inferred strategy for the subject. [Fig. 5.10]
Decision time	Average time (in seconds) spent on the screen with all the information [Fig. 5.10]
Transitions (within) [Fig. 5.10]	Average number of transitions within an alternative (inspecting different cues from the same alternative).[Fig. 5.10]

Label	Definition
Transitions (between)	Average number of transitions between alternatives (inspecting the same cue on the two alternatives). [Fig. 5.10]
Transitions (both)	Average number of transitions that included a change in both the cue and the alternative. [Fig. 5.10]
Long fixations	Average number of long fixations (500+ miliseconds in duration) [Fig. 5.10]
Long fixations (proportion)	Average proportion of long fixations (500+ miliseconds in duration) [Fig. 5.10]
Inspected cues	Average number of inspected cues. [Fig. 5.10]
Cue importance (corr. w/ contact)	Correlation between the subject's cue importance ratings and the cue importance ratings provided by the senior loan officer before the study. [Fig. 5.10]
Cue importance (hard vs. soft)	Difference between the average importance rating assigned to financial indicators and the average importance rating assigned to judgment-based cues. [Fig. 5.10]

Chapter 6

Conclusion

There's a mighty judgement coming
But I may be wrong

— Leonard Cohen predicts what this final chapter will bring

Summary:

The first part of this chapter looks back; I begin by summarizing the empirical results reported in Chapters 2 - 5 (section 6.1) and continue by discussing the results concerning process/outcome accountability against the background of existing literature on the topic (section 6.2). In particular, I note that the results reported in this dissertation do not justify a preference for process over outcome accountability and therefore appear in some conflict with the literature; I conclude by considering three candidate explanations for this apparent conflict. The second part of the chapter looks forward; I begin by summarizing a few promising directions for future research on accountability and then reflect on how the value of laboratory research on accountability could be increased (section 6.3). For example, I suggest a few steps that individual researchers and institutions could take to facilitate cumulative progress. Finally, I briefly consider the question of what the main value of laboratory experimental research on accountability is for a management system designer. I propose that, at least at this moment in time, the knowledge of *method* is more useful to a management system designer than the knowledge of *theory*.

6.1 Summary of findings

The opening chapter of this dissertation positioned accountability as an important part of the toolkit of a management system designer aiming to facilitate individual performance. A review of the literature identified several properties of accountability systems that appear most significant for this purpose. One of these is *evaluation focus*, which—according to current evidence—should typically zero in on and emphasize the *process* that the accountable individual relies on rather than the desired *outcomes*. Evaluation focus was shown to be a subject of growing academic interest and highlighted as having the potential to inform, rather than just reflect, the practice of the management system designer. The aim of Chapters 2, 3, and 4 was to make a contribution to this body of work by investigating the replicability, boundary conditions, and psychological mechanisms of the *process accountability advantage* (higher average performance under process accountability than under outcome accountability).

Throughout this dissertation, I have emphasized the importance of replication in experimental research on accountability (and in general). The topic took center stage in Chapter 2. The experiment reported in the chapter attempted to replicate two key findings originally reported by de Langhe et al. (2011). These authors investigated the effects of process and outcome accountability in *multiple-cue learning tasks* in which individuals learn to predict the criterion from a handful of cues on the basis of outcome feedback. We attempted to replicate their finding that process accountability (compared to outcome accountability) results in more accurate predictions in *elemental* multiple-cue learning tasks in which there is a linear additive relationship between the cues and the criterion *and* that this difference is mediated by the more systematic and deliberate cognitive approach of process accountable individuals (as measured by the *rational situation-specific thinking style* scale of Novak and Hoffman, 2009). The experiment also tested a potential boundary condition of the effects reported by de Langhe et al. (2011) by additionally investigating a version of the task in which some of the cues were uninformative. Finally, we used the second part of the task to investigate the effects of the two kinds of accountability on second-order probability judgments. These data presented an opportunity to replicate the finding that process accountability facilitates more accurate probability judgment than outcome accountability (Siegel-Jacobs and Yates, 1996).

Starting with the latter, our data were *not* consistent with the results of Siegel-Jacobs and Yates (1996), suggesting that probability judgment may not be reliably more accurate under process accountability than under outcome accountability. An important qualification is that the subjects in our experiment were directly held accountable for their *first-order* judgment, rather than their *second-order* (probability) judgment—a potentially consequential design difference with the experiment of Siegel-Jacobs and Yates (1996). Nevertheless, our results suggest it may be worthwhile to replicate the experiment of Siegel-Jacobs and Yates (1996). Parenthetically, there are in fact a few other reasons why replication seems worthwhile in this case. First, Siegel-Jacobs and Yates (1996) used a tournament-like approach to operationalize outcome accountability (promising a financial reward for the five most accurate subjects). It seems plausible that this manipulation incentivized risky strategies (Vandegrift and Brown, 2003) that resulted in poorer performance. The replication study should therefore use a more conventional outcome accountability manipulation. Second, a comparatively large number of subjects were excluded by Siegel-Jacobs and Yates (1996) from the outcome accountability

condition on the basis of a failed manipulation check, potentially biasing the results (see e.g. Bouwmeester et al., 2017). Finally, the influential status of Siegel-Jacobs and Yates (1996) (cited more than 300 times at the time of this writing, according to Google Scholar) and the results of Chang et al. (2017) (see section *Discussion* in Chapter 4) provide further motivation.

More reassuringly, our results were directionally consistent with those of de Langhe et al. (2011). Prediction accuracy and rational thinking style were both higher under process accountability than under outcome accountability. Furthermore, the extra measures of decision time and explicit task knowledge provided converging evidence concerning the underlying cognitive mechanism. At the same time, however, the effects we observed were substantially weaker, both in terms of statistical support and in terms of magnitude. The difference in accuracy was especially small in the novel condition in which some of the cues were uninformative. These results have clear implications for the planning of future research, although we should not lose sight of the fact that a single replication study is still just a single study.

For the effect on the rational thinking style measure, however, we can additionally examine the results from two other chapters. All results are jointly presented in Figure 6.1, which also displays the original result of de Langhe et al. (2011). The figure shows the point estimate and the classical 95% uncertainty interval for Cohen's *d*, a standardized measure of effect size. We observe considerable variation. Two of our studies indicate an effect that is directionally the same as that of de Langhe et al. (2011) but weaker, one study reveals an absence of a meaningful effect, and another study shows an effect similar in magnitude to that of de Langhe et al. (2011) — but in the opposite direction. This final result is particularly noteworthy as it is the only one that is clearly inconsistent with the original result and as it drives home the point that effects may vary not just in magnitude but also in sign.

Two of the chapters in this dissertation probed for the boundary conditions of the process accountability advantage. In Chapter 4, we investigated decision accuracy of process and outcome accountable individuals in dynamic decision environments in which decision makers face the need to combine reliable historical information and accuracy feedback. The results revealed a small but evidentiary strong performance advantage of *outcome* accountable subjects, thus qualifying the dominant view in the literature (Patil et al., 2014). As expected on the basis of preliminary results and assumptions based on previous work, this advantage was observed in the *noncompensatory* environment in which the information-frugal Take The Best heuristic was ecologically rational in the sense that its use yielded high accuracy. In addition, however, a similar result was found in the *compensatory* environment in which the Take The Best heuristic performed poorly whereas the information-intensive Weighted Additive strategy resulted in high accuracy. This finding came as a surprise because—according to previous findings and associated theorizing—this kind of environment should be favorable to process accountability. On the other hand, it is consistent with the process-tracing results which revealed a complete lack of evidence for a more systematic, deliberate, and information-intensive predecisional processing under process accountability. Overall, the results of this study demonstrate that outcome accountability can in fact support more adaptive decision making than process accountability.

Next, in Chapter 3 we examined how process and outcome accountability affect the ability to solve two types of so-called *insight problems*. This question is interesting because insight

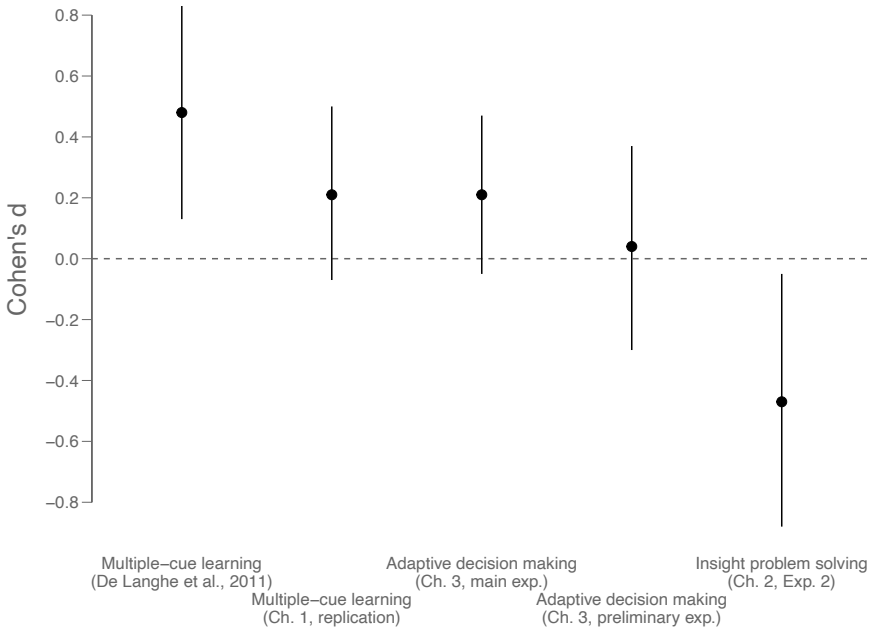


Figure 6.1: Accountability focus and rational thinking style. The x-axis displays the study, the y-axis displays standardized effect size (Cohen's d). Depicted are point estimates and classical 95% uncertainty intervals. Positive (negative) values indicate a higher average score under process (outcome) accountability.

problems provide a convenient laboratory model for an aspect of creativity that is difficult to investigate otherwise. In addition, previous research suggested that the process accountability advantage could flip in this domain. This conjecture was based on evidence linking process accountability to increased Type 2 processing and evidence suggesting that an increase in this kind of cognitive processing may be ineffective or even detrimental to performance on insight problems. Accordingly, our aim was to corroborate this theoretical account using performance and self-report data. The results of several studies, however, exposed the inadequacy of its key assumptions. First, our results contradict the assumption that process accountability reliably increases the reliance on deliberate, Type 2 thinking relative to outcome accountability (see above). Second, the correlations between measures of Type 2 processing and performance were consistently low and positive. The correlational nature of this finding means that its interpretation is not straightforward. For example, stable cognitive ability could drive variation on both these measures and performance. What is more important is that we obtained no evidence that Type 2 thinking would mediate a performance effect. Finally, the performance effects (differences in accuracy between process and outcome accountability conditions) were mixed and uncertain and the reader is directed to the Chapter 3 for a more complete exposition. Briefly, however, our studies do not suggest that process and outcome accountability have

a reliably different effect on the ability to solve matchstick arithmetic problems—the first of two tasks we studied. In contrast, we obtained some preliminary evidence that outcome accountability could be more conducive to performance on remote associate problems—the second task we studied. If shown to be robust, this result might warrant further attention.[1]

Finally, in Chapter 5 we considered the problem of decision making based on probabilistic cues but not in the context of different kinds of accountability. Whereas in Chapter 4 we investigated how process and outcome accountability affect the reliance of individuals on information-intensive and information-frugal decision strategies in response to outcome feedback, in Chapter 5 we investigated the use of these strategies among professional loan officers. This research was inspired by research on *fast-and-frugal heuristics* which suggests that information-frugal strategies such as the Take The Best heuristic can be highly accurate in real-world decision environments and perhaps even more prevalent among experienced decision makers than information-intensive decision strategies. We noted, however, that the decision environment of professional loan officers differs considerably from that of experienced decision makers investigated by previous studies, and that the characteristics of their environment may promote the use of information-intensive decision strategies. Apart from testing this possibility, the study made two additional contributions. A theoretical contribution was the development of a variant of the information-intensive Weighted Additive strategy featuring a more psychologically realistic error theory. An empirical contribution was the use of eye-tracking to provide additional information concerning the cognitive processing of loan officers. Overall, the results showed that compensatory decision strategies provided a better fit for the majority of loan officers, with the majority of the sample classified as users of the information-intensive Weighted Additive strategy. The findings from eye-tracking data resonated with this result, revealing that loan officers engaged in an exhaustive search, typically inspecting all or almost all of the information presented on the screen. Moreover, the pattern of search of most loan officers was based on the spatial arrangement of cues on the screen rather than cue importance (the latter being a search rule that is characteristic of the Take The Best strategy). These results contribute to the nascent literature on the use of fast-and-frugal heuristics among experienced decision makers.

6.2 Discussion

The following discussion focuses on the process/outcome accountability distinction, as it was central to the majority of the chapters. What would be a brief but reasonable summary of our empirical results concerning the relative differences between these two types of accountability, both those concerning performance and those concerning the underlying psychological mechanisms? Overall, the differences we found appear small, variable, and just as likely to favor outcome accountability as process accountability. From the perspective of the management system designer, these results therefore do not justify a strong preference for either process or outcome accountability. In this, they seem to be in some conflict with the literature, which has "typically championed the benefits of process accountability" (Patil et al., 2017). What explains

[1] Given the overall pattern of results from earlier studies, however, the seemingly more promising results on the remote associates task may have more to do with the fact that we studied it less.

this conflict? Let us consider three candidate answers, presented in somewhat exaggerated form.

Answer #1 denies the premise; what seems like conflict can either be explained by known moderating factors or is instead a novel finding that adds an extra building block to accountability theory. For example, the higher score of outcome accountable individuals on the rational thinking style scale found in Chapter 3 is not in conflict with de Langhe et al. (2011) but a novel finding: in insight problem solving tasks, outcome accountability is more likely to facilitate deliberate, systematic thinking than process accountability. Similarly, the higher accuracy of outcome accountable subjects in the studies of Chapter 4 means that outcome accountability yields better performance when individuals need to adaptively combine historical information and feedback from the environment. Or take the seeming inconsistency between Siegel-Jacobs and Yates (1996) and the corresponding result in Chapter 2: it reveals that the effect holds when individuals are directly accountable for their probability judgment but not if individuals are directly accountable for their primary judgment and in addition asked for a probability judgment. Similarly, our results in Chapter 3 (specifically the comparison of process accountability and no accountability) do not contradict those of Eskenazi (2015) but simply reveal the moderating role of vocabulary size. All of these results are not conflicting evidence but instead expand our knowledge of complex accountability phenomena. As Tetlock and Lerner (1999) put it, "the story to be told is one of progressive 'complexification' in which temptingly parsimonious hypotheses have been repeatedly confounded by recalcitrantly complex patterns of evidence."

Answer #2 recognizes the conflict and attributes it to the flaws of the studies in this dissertation. For example, in Chapter 1 I noted that an important component of accountability are the rewarding and punishing *consequences* for the accountable individual that may follow the evaluation. Variation concerning this component has not received much attention in the experimental literature but could in principle interact with evaluation focus. For example, the differences between process and outcome accountability may be small and unreliable when the consequences of the evaluation are relatively minor but become more pronounced as the consequences grow. Accordingly, the pattern of results observed in our research is perhaps due to accountability manipulations of insufficient "strength." This and other violations of theoretically less interesting background assumptions fully explain the results reported in this dissertation. Accordingly, paying attention to these results would simply distract us from the real gains in knowledge other studies have made.

Answer #3 says that process/outcome accountability effects are in fact highly sensitive to context and just as likely to favor outcome accountability as process accountability but that research and publishing patterns obscure this truth. After all, given the social nature of accountability and a high degree of flexibility in operationalization, we should expect these effects to vary considerably. It can be difficult to realize this, however, if independent replication efforts are missing and the reuse of tasks and measures minimal—and both of these are true for this body of work. Add the researchers' incentives and their freedom in the analysis and reporting of results (Simmons et al., 2011) plus the *file-drawer problem* (Rosenthal, 1979) to the mix and it becomes easy to see how the published record could yield a biased representation of the truth.

How credible are these three answers? In my estimation, Answer #3 is credible enough

to warrant some follow-up work and changes to research practice and direction; I provide specific recommendations in the next section. Some outside support for this conclusion comes from the recent work of Schulz-Hardt et al. (2020). These researchers attempted to replicate the finding that process accountability decreases the escalation of commitment relative to outcome accountability (Simonson and Nye, 1992). But neither the direct replication attempt nor the follow-up study (featuring "stronger" accountability manipulations) revealed any difference between the effects of these two kinds of accountability on escalation of commitment. Interestingly, the second of these two studies also featured an attempt to replicate the results of de Langhe et al. (2011). Although the full results are not reported, the authors note that "no significant effects of accountability on participants' judgments in this task were found." These findings reinforce the conclusion that the peformance advantages of process accountability are not robust.

That said, Answers #1 and #2 plausibly explain some of the "variance" as well. And accepting either one of these as the main answer can be useful to the extent that it inspires the right kind of action. For example, if one were to accept (in line with Answer #1) that the more adaptive performance of outcome accountable subjects reported in Chapter 4 constitutes a novel finding, one could follow up by conducting a replication and/or formulating and testing hypotheses concerning the underlying mechanism. Or take Answer #2, which asserts that the studies reported in this dissertation are flawed and specifically highlights the possible interaction between the evaluation focus and the "strength" of the accountability manipulations. This latter conjecture seems plausible. But it must be noted that the manipulations employed were for the most part highly similar to those of the most closely related studies. Accordingly, the critique needs to expand to include not just the manipulations but manipulations in interaction with other characteristics of the studies. Most importantly, post hoc theorizing of this kind should not be used merely as a way to selectively dismiss inconvenient results—that way a Lakatosian degeneration of a research program lies. While calling existing research on process and outcome accountability a research program in Lakatos' sense is too grandiose, the basic point stands. Dismissals of this sort are fine as long as they are generative. For example, a simple 2×2 experiment could be conducted to check if evaluation focus interacts with manipulation strength. An even more valuable solution would be to develop and validate a short and appropriately sensitive self-report instrument to tap into into this component of accountability: the subjective significance of the accountability evaluation and the associated consequences. If applied widely, such a measure would result in a body of data that could be used to test hypotheses of this sort and to guide methodological development.

6.3 Recommendations for research

Hall et al. (2017) argue that in "the history of any area of scientific inquiry, there are periods when significant opportunities present themselves, and such is the time for accountability literature." What are these opportunities? Perhaps even more importantly, *how* can researchers best exploit them? In contrast to Hall et al. (2017), we shall focus on experimental laboratory research. Let us begin with the *what* and end on the *how*.

In relation to the process/outcome accountability distinction, one aspect of Answer #3 in the

previous section is worth highlighting because it leads to my first recommendation. Recall that the first component of Answer #3 is the argument that process/outcome accountability effects are highly variable. The basis of the argument are two sources of heterogeneity: flexibility in operationalization/implementation and the open-ended and social nature of accountability manipulations. But note that if we could assume that the published record provides an unbiased reflection of all the research that was conducted, the combination of these heterogeneity sources and relative uniformity in the main result would provide some cause for celebration. For it would hint that the distinction between process and outcome acccountability has considerable explanatory power—one that does not depend strongly on, for example, choosing *this* rather than *that* operationalization of outcome accountability. Alas, this assumption is not tenable (the second component of Answer #3), which suggests that a useful next step for future research would be a literature review and a meta-analysis that would be based on a serious effort to uncover and assess relevant research on process and outcome accountability that has not been published. Such a project would help evaluate the validity of Answer #3. Relatedly, statistical approaches that aim to assess the degree of publication bias could also prove helpful.

Another next step addressing the issue of heterogeneity of accountability effects would be to conduct a larger study (involving a number of labs) with the aim of explicitly investigating the magnitude of this issue. Recently, a "Many Labs" collaboration effort explored this very question for a set of 28 classic and contemporary findings in psychology (Klein et al., 2018). Each finding was tested in about half of the 125 samples (spanning more than 15 000 subjects in total) and heterogeneity across samples, cultures, etc. was assessed in addition to the basic average effect. This body of results would provide a useful source of benchmarks if a project of this sort would be carried out in the accountability literature.

Of course, projects of this kind are resource-heavy. This would be even more true for typical accountability studies, which do not lend themselves to quick administration requiring little human involvement. Accordingly, in comparison to my first suggestion, this one is more debatable and one may reasonably conclude that such a large investment of resources is not warranted. Especially if, as Patil et al. (2014) suggest, "the effects of process and outcome accountability hinge less on the process–outcome distinction than they do on the social-identity signals that accountability sends to employees about how managers view them." But for the researchers who *do* believe in the usefulness of the process-outcome distinction such a project would be worthwhile.

Beyond the proces/outcome accountability distinction, the search for useful accountability knowledge continues. An important direction for future research concerns the negative consequences of accountability systems, especially those that are most likely and consequential yet are also easy to miss or even run counter to intuition. In particular, the imposition of a certain accountability system could have an unrecognized *signaling* component that would lead to undesired behavior or depress desired behavior. For example, accountability systems can provide information on the assumptions the management system designer is making about the accountable individuals or information about the prevalence of a certain behavior. As Patil et al. (2014) note, these kind of signals could be consequential. Accordingly, an important type of research question is how to introduce, frame, and position accountability systems so that their benefits are maximized while the downsides are avoided.

Another kind of undesirable behavior that can easily arise under accountability is *defensive decision making*, which occurs when "professionals opt for the second-best option rather than (what they believe to be) the best option for their organization or client in order to protect themselves from potential negative consequences in the future" (Artinger et al., 2019). The phenomenon has received attention in medicine, where physicians frequently order unneccessary diagnostic tests and engage in other behaviors that may fulfill a protective function in case of undesired patient outcomes but are, on the whole, suboptimal (Studdert, 2005). Among managers, Artinger et al. (2019) investigated the prevalence of the phenomenon in a large public administration in Germany and reported that one out of four of the most important decisions in the previous year featured defensive decision making. An important goal of accountability research is discovering how the presence and intensity of such "side effects" can be minimized without losing the beneficial effects of accountability. Of course, this goal extends beyond decision making; the tension between accountability and creativity, for example, remains an important problem.

Apart from investigating the deleterious effects of accountability and ways of attenuating or even eliminating such effects, another important research direction involves determining the critical characteristics of *positive* effects of accountability that have been observed in laboratory settings. This means answering the question: Which features of an accountability manipulation must the management system designer replicate in order to obtain the reported benefits? For example, the role of *information (a)symmetry* between the accountable individual and the evaluating audience has, to the best of my knowledge, not been experimentally investigated yet. In many laboratory settings in which positive effects of accountability have been demonstrated this factor favors the evaluating audience (often the experimenter or their research assistant), which has access to all relevant information. In contrast, in many real-world settings there is at least some informational asymmetry running in the other direction, with the potential to attenuate accountability benefits.

This brings us to the question of *how*. Apart from sharpening the ability to zero in on the right kind of questions and generate useful, insightful, and precise theories, what can we do to maximize the value of laboratory research on accountability (and more generally)? Is the current *modus operandi* of laboratory research ideally suited to this aim or are there steps that would enable faster cumulative progress while also being feasible? My basic position is that, in answering this question, the appropriate level of analysis is not an individual research project or paper but a larger body of work—a stream of literature, a research program. Actions that might seem limited in value in the context of a single research project can be an important driver of progress if they become common.

Some of the recommendations I put forward in Chapter 1 seem to belong in this category. For example, the recommendation to openly share data and materials may seem more trouble than it is worth in the context of a given research project. But if applied widely, the practice facilitates cumulative progress by making easier the conduct of replication studies and the testing of novel hypotheses. Perhaps most importantly, it enables other researchers to carefully scrutinize a given finding or claim. The recommendation to conduct replication studies provides another demonstration. Conducting a replication study consumes resources that could be spent on an original study with the potential to yield exciting novel findings. And a desire for exploration and novel findings is perfectly justifiable. But if indulged fully, it results in a lack of replication

studies and a literature that is poorer for it.

Next, I believe that increased collaboration and joint research efforts would go a long way toward facilitating cumulative progress. Surely the value of laboratory research is not maximized by a series of independent, small-*N* two-by-twos. This becomes increasingly true as time goes by and more and more low-hanging fruit are picked and squeezed into cider. Accordingly, the reach of laboratory research on accountability could be extended by larger-scale collaborative efforts that can tackle more ambitious questions and provide stronger evidence than a series of smaller, independent studies. Recent initiatives of this sort in psychology, such as the "Many Labs" investigations (e.g. Klein et al., 2014, 2018; Ebersole et al., 2016) and the "Psychological Science Accelerator" project (Moshontz et al., 2018), are therefore highly welcome developments and can serve as models. Although such projects are costly, they provide—at least for *some* research questions—a more efficient approach over the long run. Moreover, it is not only about truly large-scale projects that feature tens of different labs or groups of researchers and thousands of participants; I believe important gains could also be made by collaborations on a smaller scale. For example, an independent test of the same hypothesis by two (groups of) researchers could already be highly informative.

Although there is a lot that individual researchers can do, there are limits. It is not just that time and energy are limited; career-wise, rigor can be counterproductive if improper incentives are in place. Smaldino and McElreath (2016) show how a process of natural selection—in an environment modeled after the current academic environment—can lead to a propagation of poor research practices and a high false discovery rates. Importantly, this process "requires no conscious strategizing nor cheating on the part of researchers." Universities, funding agencies, and academic journals therefore play an essential role; to ensure optimal progress in (accountability) research, the accountability of (accountability) researchers needs to be carefully considered and properly implemented. For example, journals could contribute by incorporating a diverse set of publishing options (traditional reports, registered reports, brief reports, theory papers, etc.) and other mechanisms that would counter publication bias. And universities and funding agencies can contribute by devising appropriate evaluation systems that go beyond publication and citation counts. In business, the recognition that traditional financial performance measures are not enough led to the development of a *balanced scorecard*, which features a richer set of lagging and leading indicators (Kaplan and Norton, 1992). In science, the development of appropriate "scientific balanced scorecards" by the relevant institutions may do more than anything else to facilitate cumulative progress. In science, as in business, accountability reigns supreme.

6.4 Recommendations for practice

Where does all this leave our management system designer? Thinking clearly about accountability is essential whenever complex human behavior needs to be predicted and influenced, and experimental laboratory research can assist in this task by highlighting ideas that have been subjected to some conceptual and empirical scrutiny. At the same time, there are reasons to be skeptical of the output of the evolutionary process that operates in the current academic environment (Smaldino and McElreath, 2016). Furthermore, the social contingency model

(Tetlock and Lerner, 1999) and the results reported in this dissertation suggest accountability effects can be highly sensitive to implementation and context. Accordingly, the current academic literature on accountability is better thought of as a source of ideas rather than a repository of established knowledge. There is also the fact that current theorizing and empirical results offer only limited guidance in setting up a real-world accountability system; ambiguity arises in questions of *implementation* (which of the many ways of introducing process accountability should I choose?) but also in questions of *impact* (just how much money will this save us?). In decision making, being certain about the direction of an effect sometimes means little if paired with great uncertainty about the magnitude.

In light of these observations, the greater contribution of experimental laboratory research to the practice of the management system designer may in fact be not *theory* but *method*. On this view, the management system designer would be well-advised to learn from the body of knowledge on experimental design and learning from data (or collaborate with academic researchers) and then use this knowledge when setting up and assessing the results of various accountability initiatives. Just as the laboratory researcher attempts to capture the world in the lab, so could the management system designer attempt to capture the lab in the world. Though such efforts to rigorously evaluate accountability initiatives may not come easy or cheap, in the long run they would likely lead to better outcomes than blindly trusting intuition, theory, or even ideology (Tetlock et al., 2013).

This view resonates with Campbell's vision of the *experimenting society* (Campbell, 1969, 1991). Expressing "great pessimism in the relevance of our theory for applied settings," Campbell suggested that the body of knowledge on experimentation and learning from data, "rather than our disciplinary knowledge, is what we social scientists have to offer to the social policy process." At this moment in time, a similar conclusion seems to apply to the case of a management system designer looking to learn from laboratory experimental research on accountability.

References

Abdel-Khalik, A. R. and El-Sheshai, K. M. (1980). Information Choice and Utilization in an Experiment on Default Prediction. *Journal of Accounting Research*, 18(2):325.

Aleksovska, M., Schillemans, T., and Grimmelikhuijsen, S. (2019). Lessons from five decades of experimental and behavioral research on accountability: A systematic literature review. *Journal of Behavioral Public Administration*, 2(2).

Alexander, J. K., Hillier, A., Smith, R. M., Tivarus, M. E., and Beversdorf, D. Q. (2007). Beta-adrenergic modulation of cognitive flexibility during stress. *Journal of cognitive neuroscience*, 19(3):468–478.

Amabile, T. (1998). How to Kill Creativity. *Harvard Business Review*, (September–October 1998).

Amer, T., Campbell, K. L., and Hasher, L. (2016). Cognitive Control As a Double-Edged Sword. *Trends in Cognitive Sciences*, 20(12):905–915.

Andersson, P. (2004). Does experience matter in lending? A process-tracing study on experienced loan officers' and novices' decision behavior. *Journal of Economic Psychology*, 25(4):471–492.

Ariely, D. (2008). *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. Harper, New York, NY, 1st ed edition.

Arkes, H. R., Gigerenzer, G., and Hertwig, R. (2016). How bad is incoherence? *Decision*, 3(1):20–39.

Artinger, F., Petersen, M., Gigerenzer, G., and Weibler, J. (2015). Heuristics as adaptive decision strategies in management. *Journal of Organizational Behavior*, 36(S1):S33–S52.

Artinger, F. M., Artinger, S., and Gigerenzer, G. (2019). C. Y. A.: Frequency and causes of defensive decisions in public administration. *Business Research*, 12(1):9–25.

Ball, L. and Stevens, A. (2009). Evidence for a verbally-based analytic component to insight problem solving. In *Proceedings of the Cognitive Science Society*, volume 31.

Ball, L. J., Marsh, J. E., Litchfield, D., Cook, R. L., and Booth, N. (2015). When distraction helps: Evidence that concurrent articulation and irrelevant speech can facilitate insight problem solving. *Thinking & Reasoning*, 21(1):76–96.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White,

C. N., De Boeck, P., and Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, 115(11):2607–2612.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, 67(1).

Beaulieu, P. R. (1994). Commercial lenders' use of accounting information in interaction with source credibility. *Contemporary Accounting Research; Toronto*, 10(2):557.

Bernstein, E. and Waber, B. (2019). The Truth About Open Offices. *Harvard Business Review*, (November–December 2019).

Biggs, S. F., Bedard, J. C., Gaber, B. G., and Linsmeier, T. J. (1985). The effects of task size and similarity on the decision behavior of bank loan officers. *Management Science*, 31(8):970–987.

Bol, J. C. (2011). The Determinants and Performance Effects of Managers' Performance Evaluation Biases. *The Accounting Review*, 86(5):1549–1575.

Bouwmeester, S., Verkoeijen, P. P. J. L., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., Chmura, T. G. H., Cornelissen, G., Døssing, F. S., Espín, A. M., Evans, A. M., Ferreira-Santos, F., Fiedler, S., Flegr, J., Ghaffari, M., Glöckner, A., Goeschl, T., Guo, L., Hauser, O. P., Hernan-Gonzalez, R., Herrero, A., Horne, Z., Houdek, P., Johannesson, M., Koppel, L., Kujal, P., Laine, T., Lohse, J., Martins, E. C., Mauro, C., Mischkowski, D., Mukherjee, S., Myrseth, K. O. R., Navarro-Martínez, D., Neal, T. M. S., Novakova, J., Pagà, R., Paiva, T. O., Palfi, B., Piovesan, M., Rahal, R.-M., Salomon, E., Srinivasan, N., Srivastava, A., Szaszi, B., Szollosi, A., ø. Thor, K., Tinghög, G., Trueblood, J. S., Van Bavel, J. J., van 't Veer, A. E., Västfjäll, D., Warner, M., Wengström, E., Wills, J., and Wollbrant, C. E. (2017). Registered Replication Report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, 12(3):527–542.

Bowden, E., Jungbeeman, M., Fleck, J., and Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, 9(7):322–328.

Bowden, E. M. and Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods*, 35(4):634–639.

Breaux, D. M., Munyon, T. P., Hochwarter, W. A., and Ferris, G. R. (2009). Politics as a Moderator of the Accountability—Job Satisfaction Relationship: Evidence Across Three Studies. *Journal of Management*, 35(2):307–326.

Bretz, R. D., Milkovich, G. T., and Read, W. (1992). The Current State of Performance Appraisal Research and Practice: Concerns, Directions, and Implications. *Journal of Management*, 18(2):321–352.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):3.

Brighton, H. (2006). Robust Inference with Simple Cognitive Models. In *AAAI Spring Symposium: Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems*, page 6.

Bröder, A. (2000). Assessing the empirical validity of the "Take-the-best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5):1332–1346.

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(4):611–625.

Bröder, A. and Schiffer, S. (2003). Bayesian strategy assessment in multi-attribute decision making: Strategy Assessment in Decision Making. *Journal of Behavioral Decision Making*, 16(3):193–213.

Bröder, A. and Schiffer, S. (2006). Adaptive flexibility and maladaptive routines in selecting fast and frugal decision strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(4):904–918.

Brooks, S. P. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Brown, C. L. (1999). "Do the Right Thing:" Diverging Effects of Accountability in a Managerial Context. *Marketing Science*, 18(3):230–246.

Brtek, M. D. and Motowidlo, S. J. (2002). Effects of procedure and outcome accountability on interview validity. *Journal of Applied Psychology*, 87(1):185–191.

Bürkner, P.-C. (2017). Brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1).

Bürkner, P.-C. and Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):25.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637.

Camerer, C. F. and Hogarth, R. M. (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty*, 19(1/3):7–42.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, 24(4):409–429.

Campbell, D. T. (1991). Methods for the Experimenting Society. *Evaluation practice*, page 38.

Casey, C. J. (1980). Variation in Accounting Information Load: The Effect on Loan Officers' Predictions of Bankruptcy. *The Accounting Review*, 55(1):15.

Catasús, B. and Gröjer, J.-E. (2003). Intangibles and credit decisions: Results from an experiment. *European Accounting Review*, 12(2):327–355.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, 49(3):609–610.

Chang, L. J., Cheng, M. M., and Trotman, K. T. (2013). The effect of outcome and process accountability on customer–supplier negotiations. *Accounting, Organizations and Society*, 38(2):93–107.

Chang, W., Atanasov, P., Patil, S., Mellers, B. A., and Tetlock, P. E. (2017). Accountability and adaptive performance under uncertainty: A long-term view. *Judgment and Decision Making*, page 17.

Chein, J. M. and Weisberg, R. W. (2014). Working memory and insight in verbal problems: Analysis of compound remote associates. *Memory & Cognition*, 42(1):67–83.

Chein, J. M., Weisberg, R. W., Streeter, N. L., and Kwok, S. (2010). Working memory and insight in the nine-dot problem. *Memory & Cognition*, 38(7):883–892.

Choi, J. W., Hecht, G. W., and Tayler, W. B. (2012). Lost in Translation: The Effects of Incentive Compensation on Strategy Surrogation. *The Accounting Review*, 87(4):1135–1163.

Chu, Y. and MacGregor, J. N. (2011). Human Performance on Insight Problem Solving: A Review. *The Journal of Problem Solving*, 3(2).

Chuderski, A. and Jastrzębski, J. (2017). Working memory facilitates insight instead of hindering it: Comment on DeCaro, Van Stockum, and Wieth (2016). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(12):1993–2004.

Chuderski, A. and Jastrzębski, J. (2018). Much ado about aha!: Insight problem solving is strongly related to working memory capacity and reasoning ability. *Journal of Experimental Psychology: General*, 147(2):257–281.

Cohen, J. D. (2017). Cognitive Control: Core Constructs and Current Considerations. In Egner, T., editor, *The Wiley Handbook of Cognitive Control*, pages 1–28. John Wiley & Sons, Ltd, Chichester, UK.

Czerlinski, J., Gigerenzer, G., and Goldstein, D. G. (1999). How good are simple heuristics? In *Simple Heuristics That Make Us Smart*, pages 97–118. Oxford University Press, London.

Dalla Via, N., Perego, P., and van Rinsum, M. (2018). How accountability type influences information search processes and decision quality. *Accounting, Organizations and Society*.

Davis, W. D., Mero, N., and Goodman, J. M. (2007). The Interactive Effects of Goal Orientation and Accountability on Task Performance. *Human Performance*, 20(1):1–21.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, pages 571–582.

De Dreu, C. K. W., Beersma, B., Stroebe, K., and Euwema, M. C. (2006). Motivated information processing, strategic choice, and the quality of negotiated agreement. *Journal of Personality and Social Psychology*, 90(6):927–943.

de Dreu, C. K. W., Koole, S. L., and Steinel, W. (2000). Unfixing the fixed pie: A motivated information-processing approach to integrative negotiation. *Journal of Personality and Social Psychology*, 79(6):975–987.

de Langhe, B., van Osselaer, S. M., and Wierenga, B. (2011). The effects of process and outcome accountability on judgment process and performance. *Organizational Behavior and Human Decision Processes*, 115(2):238–252.

DeCaro, M. S., Van Stockum, C. A., and Wieth, M. B. (2016). When higher working memory capacity hinders insight. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(1):39–49.

DeCaro, M. S., Van Stockum, C. A., and Wieth, M. B. (2017). The relationship between working memory and insight depends on moderators: Reply to Chuderski and Jastrzêbski (2017). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(12):2005–2010.

DeCaro, M. S. and Van Stockum, C. A. J. (2017). Ego depletion improves insight. *Thinking & Reasoning*, 0(0):1–29.

Demeré, B. W., Sedatole, K. L., and Woods, A. (2018). The Role of Calibration Committees in Subjective Performance Evaluation Systems. *Management Science*, 65(4):1562–1585.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal Versus Naive Diversification: How Inefficient is the 1/ *N* Portfolio Strategy? *Review of Financial Studies*, 22(5):1915–1953.

Dhami, M. K. and Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making*, 14(2):141–168.

Dickert, S., Rubaltelli, E., and Slovic, P. (2012). Response Mode, Compatibility, and Dual-processes in the Evaluation of Simple Gambles: An eye-tracking investigation: (519682015-052).

Dieckmann, A. and Rieskamp, J. (2007). The influence of information redundancy on probabilistic inferences. *Memory & Cognition*, 35(7):1801–1813.

Doney, P. M. and Armstrong, G. M. (1996). Effects of accountability on symbolic information search and information analysis by organizational buyers. *Journal of the Academy of Marketing Science*, 24(1):57–65.

Dummel, S., Rummel, J., and Voss, A. (2016). Additional information is not ignored: New evidence for information integration and inhibition in take-the-best decisions. *Acta Psychologica*, 163:167–184.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., Davis, W. E., Devos, T., Fletcher, M. M., German, K., Grahe, J. E., Hermann, A. D., Hicks, J. A., Honeycutt, N., Humphrey, B., Janus, M., Johnson, D. J., Joy-Gaba, J. A.,

Juzeler, H., Keres, A., Kinney, D., Kirshenbaum, J., Klein, R. A., Lucas, R. E., Lustgraaf, C. J. N., Martin, D., Menon, M., Metzger, M., Moloney, J. M., Morse, P. J., Prislin, R., Razza, T., Re, D. E., Rule, N. O., Sacco, D. F., Sauerberger, K., Shrider, E., Shultz, M., Siemsen, C., Sobocko, K., Weylin Sternglanz, R., Summerville, A., Tskhay, K. O., van Allen, Z., Vaughn, L. A., Walker, R. J., Weinberg, A., Wilson, J. P., Wirth, J. H., Wortman, J., and Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67:68–82.

Einhorn, H. J. and Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, 13(2):171–192.

Eskenazi, P. (2015). *The Accountable Animal*. PhD thesis, Erasmus University Rotterdam, Rotterdam.

Evans, J. S. B. T. and Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, 8(3):223–241.

Fehrenbacher, D. D., Kaplan, S. E., and Moulang, C. (2020). The role of accountability in reducing the impact of affective reactions on capital budgeting decisions. *Management Accounting Research*, 47:100650.

Fleck, J. I. (2008). Working memory demands in insight versus analytic problem solving. *European Journal of Cognitive Psychology*, 20(1):139–176.

Fox, F. V. and Staw, B. M. (1979). The Trapped Administrator: Effects of Job Insecurity and Policy Resistance Upon Commitment to a Course of Action. *Administrative Science Quarterly*, 24(3):449.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, 19(4):25–42.

Frith, C. D. and Frith, U. (2006). The Neural Basis of Mentalizing. *Neuron*, 50(4):531–534.

Garbers, Y. and Konradt, U. (2014). The effect of financial incentives on performance: A quantitative review of individual and team-based financial incentives. *Journal of Occupational and Organizational Psychology*, 87(1):102–137.

Garcia-Retamero, R. and Dhami, M. K. (2009). Take-the-best in expert-novice decision strategies for residential burglary. *Psychonomic Bulletin & Review*, 16(1):163–169.

Gelman, A. (2018). Don't characterize replications as successes or failures. *Behavioral and Brain Sciences.*, 41.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543.

Gigerenzer, G. (2008). Why Heuristics Work. *Perspectives on Psychological Science*, 3(1):20–29.

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, 1(2):198–218.

Gigerenzer, G. and Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, 1(1):107–143.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, 62(1):451–482.

Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669.

Gigerenzer, G. and Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In *Simple Heuristics That Make Us Smart*, pages 75–95. Oxford University Press.

Gigerenzer, G. and Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making*, 6(1):22.

Gigerenzer, G. and Todd, P. M., editors (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press, USA.

Gilhooly, K. J., Fioratou, E., and Henretty, N. (2010). Verbalization and problem solving: Insight and spatial factors. *British Journal of Psychology*, 101(1):81–93.

Glöckner, A. (2009). Investigating intuitive and deliberate processes statistically: The multiple-measure maximum likelihood strategy classification method. *Judgment and Decision Making*, 4(3):15.

Goldstein, D. G. and Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109(1):75–90.

Greenwald, A. G. (1975). Consequences of Prejudice Against the Null Hypothesis. *Psychological Bulletin*, 82(1):1–20.

Hall, A. T. and Ferris, G. R. (2011). Accountability and Extra-Role Behavior. *Employee Responsibilities and Rights Journal*, 23(2):131–144.

Hall, A. T., Frink, D. D., and Buckley, M. R. (2017). An accountability account: A review and synthesis of the theoretical and empirical research on felt accountability: Accountability. *Journal of Organizational Behavior*, 38(2):204–224.

Hall, A. T., Royle, M. T., Brymer, R. A., Perrewé, P. L., Ferris, G. R., and Hochwarter, W. A. (2006). Relationships between felt accountability as a stressor and strain reactions: The neutralizing role of autonomy across two studies. *Journal of Occupational Health Psychology*, 11(1):87–99.

Häusser, J. A., Frisch, J. U., Wanzel, S., and Schulz-Hardt, S. (2017). Effects of Process and Outcome Accountability on Idea Generation. *Experimental Psychology*, 64(4):262–272.

Heck, D. W., Hilbig, B. E., and Moshagen, M. (2017). From information processing to decisions: Formalizing and comparing psychologically plausible choice models. *Cognitive Psychology*, 96:26–40.

Hilbig, B. E. and Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review*, 21(6):1431–1443.

Hochwarter, W. A., Ferris, G. R., Gavin, M. B., Perrewé, P. L., Hall, A. T., and Frink, D. D. (2007). Political skill as neutralizer of felt accountability—job tension effects on job performance ratings: A longitudinal investigation. *Organizational Behavior and Human Decision Processes*, 102(2):226–239.

Hochwarter, W. A., Kacmar, C., and Ferris, G. R. (2003). Accountability at work: An examination of antecedents and consequences.

Hochwarter, W. A., Perrewé, P. L., Hall, A. T., and Ferris, G. R. (2005). Negative Affectivity as a Moderator of the Form and Magnitude of the Relationship between Felt Accountability and Job Tension. *Journal of Organizational Behavior*, 26(5):517–534.

Hoffman, V. B. and Patton, J. M. (1997). Accountability, the Dilution Effect, and Conservatism in Auditors' Fraud Judgments. *Journal of Accounting Research*, 35(2):227.

Hoffrage, U., HafenbrÃ¤dl, S., and Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in Psychology*, 6.

Hogarth, R. (2012). When Simple Is Hard to Accept. In *Ecological Rationality: Intelligence in the World*. Oxford University Press.

Hogarth, R. M. and Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review*, 114(3):733–758.

Hogarth, R. M., Lejarraga, T., and Soyer, E. (2015). The Two Settings of Kind and Wicked Learning Environments. *Current Directions in Psychological Science*, 24(5):379–385.

Holmqvist, K., Nystrom, M., Andersson, R., Dewhurst, R., Jarodzka, H., and van de Weijer, J. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, Oxford, first published in paperback edition.

Horstmann, N., Ahlgrimm, A., and Glöckner, A. (2009). How Distinct are Intuition and Deliberation? An Eye-Tracking Analysis of Instruction-Induced Decision Modes. *Judgment and Decision Making*, 4(5):335–354.

Hsee, C. K. (1996). The Evaluability Hypothesis: An Explanation for Preference Reversals between Joint and Separate Evaluations of Alternatives. *Organizational Behavior and Human Decision Processes*, 67(3):247–257.

Hüffmeier, J., Mazei, J., and Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, 66:81–92.

Janis, I. L. and Mann, L. (1977). *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment*. Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment. Free Press, New York, NY, US.

Jarosz, A. F., Colflesh, G. J., and Wiley, J. (2012). Uncorking the muse: Alcohol intoxication facilitates creative problem solving. *Consciousness and Cognition*, 21(1):487–493.

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

Kahn, B. E. and Baron, J. (1995). An Exploratory Study of Choice Rules Favored for High-Stakes Decisions. *Journal of Consumer Psychology*, 4(4):305–328.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kämmer, J. E., Gaissmaier, W., Expt, U. C., et al. (2013). The environment matters: Comparing individuals and dyads in their adaptive use of decision strategies. *Judgment and Decision Making*, 8(3):299.

Kaplan, R. S. and Norton, D. P. (1992). The Balanced Scorecard: Measures That Drive Performance. *Harvard Business Review*, (January-February 1992).

Kareev, Y. and Avrahami, J. (2007). Choosing between Adaptive Agents: Some Unexpected Implications of Level of Scrutiny. *Psychological Science*, 18(7):636–641.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Kennedy, J. (1993). Debiasing Audit Judgment with Accountability: A Framework and Experimental Results. *Journal of Accounting Research*, 31(2):231.

Kim, S. and Trotman, K. T. (2015). The comparative effect of process and outcome accountability in enhancing professional scepticism. *Accounting & Finance*, 55(4):1015–1040.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M.-S., Joy-Gaba, J. A., Barry Kappes, H., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van 't Veer, A. E., Ann Vaughn, L., Vranka, M., Wichman, A. L., Woodzicka, J. A., and Nosek, B. A. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology*, 45(3):142–152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., Rédei, A. C., Cai, H., Cambier, F., Cantarero, K., Carmichael, C. L., Ceric, F., Chandler, J., Chang, J.-H., Chatard, A., Chen, E. E., Cheong, W., Cicero, D. C., Coen, S., Coleman, J. A., Collisson, B., Conway, M. A., Corker, K. S., Curran, P. G., Cushman, F., Dagona, Z. K., Dalgar, I., Dalla Rosa, A., Davis, W. E., de Bruijn, M., De Schutter, L., Devos, T., de Vries, M., Doğulu, C., Dozo, N., Dukes, K. N., Dunham, Y., Durrheim, K., Ebersole, C. R., Edlund, J. E., Eller, A., English, A. S., Finck, C., Frankowska, N., Freyre, M.-Á., Friedman, M., Galliani, E. M., Gandi, J. C., Ghoshal, T., Giessner, S. R., Gill, T., Gnambs, T., Gómez, Á., González, R., Graham, J., Grahe, J. E., Grahek, I., Green, E. G. T., Hai, K., Haigh, M., Haines, E. L., Hall, M. P., Heffernan, M. E., Hicks, J. A., Houdek, P., Huntsinger, J. R., Huynh, H. P., IJzerman, H., Inbar, Y., Innes-Ker, Å. H., Jiménez-Leal, W., John, M.-S., Joy-Gaba, J. A., Kamiloğlu, R. G., Kappes, H. B., Karabati, S., Karick, H., Keller, V. N., Kende, A., Kervyn, N., Knežević, G., Kovacs, C., Krueger, L. E., Kurapov, G., Kurtz, J., Lakens, D., Lazarević, L. B., Levitan, C. A., Lewis, N. A.,

Lins, S., Lipsey, N. P., Losee, J. E., Maassen, E., Maitner, A. T., Malingumu, W., Mallett, R. K., Marotta, S. A., Meedović, J., Mena-Pacheco, F., Milfont, T. L., Morris, W. L., Murphy, S. C., Myachykov, A., Neave, N., Neijenhuijs, K., Nelson, A. J., Neto, F., Lee Nichols, A., Ocampo, A., O'Donnell, S. L., Oikawa, H., Oikawa, M., Ong, E., Orosz, G., Osowiecka, M., Packard, G., Pérez-Sánchez, R., Petrović, B., Pilati, R., Pinter, B., Podesta, L., Pogge, G., Pollmann, M. M. H., Rutchick, A. M., Saavedra, P., Saeri, A. K., Salomon, E., Schmidt, K., Schönbrodt, F. D., Sekerdej, M. B., Sirlopú, D., Skorinko, J. L. M., Smith, M. A., Smith-Castro, V., Smolders, K. C. H. J., Sobkow, A., Sowden, W., Spachtholz, P., Srivastava, M., Steiner, T. G., Stouten, J., Street, C. N. H., Sundfelt, O. K., Szeto, S., Szumowska, E., Tang, A. C. W., Tanzer, N., Tear, M. J., Theriault, J., Thomae, M., Torres, D., Traczyk, J., Tybur, J. M., Ujhelyi, A., van Aert, R. C. M., van Assen, M. A. L. M., van der Hulst, M., van Lange, P. A. M., van 't Veer, A. E., Vásquez- Echeverría, A., Ann Vaughn, L., Vázquez, A., Vega, L. D., Verniers, C., Verschoor, M., Voermans, I. P. J., Vranka, M. A., Welch, C., Wichman, A. L., Williams, L. A., Wood, M., Woodzicka, J. A., Wronska, M. K., Young, L., Zelenski, J. M., Zhijia, Z., and Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490.

Klimoski, R. and Inks, L. (1990). Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes*, 45(2):194–208.

Knoblich, G., Ohlsson, S., Haider, H., and Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(6):1534–1555.

Knoblich, G., Ohlsson, S., and Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition*, 29(7):1000–1009.

Lahey, J. N. and Oxley, D. (2016). The Power of Eye Tracking in Economics Experiments. *American Economic Review*, 106(5):309–313.

Lavric, A., Forstmeier, S., and Rippon, G. (2000). Differences in working memory involvement in analytical and creative tasks: An ERP study. *NeuroReport*, 11(8):1613–1618.

LeBel, E. P., Vanpaemel, W., Cheung, I., and Campbell, L. (2019). A Brief Guide to Evaluate Replications. *Meta-Psychology*, 3.

Lee, H., Herr, P. M., Kardes, F. R., and Kim, C. (1999). Motivated Search. *Journal of Business Research*, 45(1):75–88.

Lee, M. D. (2016). Bayesian outcome-based strategy classification. *Behavior Research Methods*, 48(1):29–41.

Lee, M. D. (2019). A simple and flexible Bayesian method for inferring step changes in cognition. *Behavior Research Methods*, 51(2):948–960.

Lee, M. D., Blanco, G., and Bo, N. (2017). Testing take-the-best in new and changing environments. *Behavior Research Methods*, 49(4):1420–1431.

Lee, M. D., Gluck, K. A., and Walsh, M. M. (2019). Understanding the complexity of simple decisions: Modeling multiple behaviors and switching strategies. *Decision*.

Lee, M. D. and Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Lemhöfer, K. and Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, 44(2):325–343.

Lerner, J. S. and Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125(2):255–275.

Libby, R. (1975). Accounting Ratios and the Prediction of Failure: Some Behavioral Evidence. *Journal of Accounting Research*, 13(1):150.

Libby, T., Salterio, S. E., and Webb, A. (2004). The Balanced Scorecard: The Effects of Assurance and Process Accountability on Managerial Judgment. *The Accounting Review*, 79(4):1075–1094.

Lilienfeld, S. O., Ammirati, R., and Landfield, K. (2009). Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare? *Perspectives on Psychological Science*, 4(4):390–398.

Lohse, G. L. and Johnson, E. J. (1996). A Comparison of Two Process Tracing Methods for Choice Tasks. *Organizational Behavior and Human Decision Processes*, 68(1):28–43.

Luan, S., Reb, J., and Gigerenzer, G. (2019). Ecological Rationality: Fast-and-Frugal Heuristics for Managerial Decision Making under Uncertainty. *Academy of Management Journal*.

Lynch, E. J. and Andiola, L. M. (2018). If Eyes are the Window to Our Soul, What Role does Eye Tracking Play in Accounting Research? *Behavioral Research in Accounting*, page 53.

MacGregor, J. N., Ormerod, T. C., and Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):176–201.

Mackey, J. D., Brees, J. R., McAllister, C. P., Zorn, M. L., Martinko, M. J., and Harvey, P. (2018). Victim and Culprit? The Effects of Entitlement and Felt Accountability on Perceptions of Abusive Supervision and Perpetration of Workplace Bullying. *Journal of Business Ethics*, 153(3):659–673.

Marewski, J. N., Schooler, L. J., and Gigerenzer, G. (2010). Five Principles for Studying People's Use of Heuristics: Five Principles for Studying People's Use of Heuristics. *Acta Psychologica Sinica*, 42(1):72–87.

Markman, K. D. and Tetlock, P. E. (2000). Accountability and Close-Call Counterfactuals: The Loser Who Nearly Won and the Winner Who Nearly Lost. *Personality and Social Psychology Bulletin*, 26(10):1213–1224.

Marteau, T. M. and Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology*, 31(3):301–306.

Martindale, C. and Greenough, J. (1973). The Differential Effect of Increased Arousal on Creative and Intellectual Performance. *The Journal of Genetic Psychology*, 123(2):329–335.

Mata, R., von Helversen, B., and Rieskamp, J. (2010). Learning to choose: Cognitive aging and strategy selection learning in decision making. *Psychology and Aging*, 25(2):299–309.

Matton, N., Raufaste, É., and Vautier, S. (2013). External validity of individual differences in multiple cue probability learning: The case of pilot training. *Judgment and Decision Making*, 8(5):14.

McGraw, K. O. and Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111(2):361–365.

Mednick, S. (1962). The associative basis of the creative process. *Psychological review*, 69(3):220.

Meißner, M. and Oll, J. (2019). The Promise of Eye-Tracking Methodology in Organizational Research: A Taxonomy, Review, and Future Avenues. *Organizational Research Methods*, 22(2):590–617.

Merchant, K. A. and Otley, D. T. (2006). A Review of the Literature on Control and Accountability. In *Handbooks of Management Accounting Research*, volume 2, pages 785–802. Elsevier.

Moser, K., Wolff, H.-G., and Kraft, A. (2013). The de-escalation of commitment: Predecisional accountability and cognitive processes. *Journal of Applied Social Psychology*, 43(2):363–376.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., Álvarez Solas, S., Ansari, D., Awlia, D., Baskin, E., Batres, C., Borras-Guevara, M. L., Brick, C., Chandel, P., Chatard, A., Chopik, W. J., Clarance, D., Coles, N. A., Corker, K. S., Dixson, B. J. W., Dranseika, V., Dunham, Y., Fox, N. W., Gardiner, G., Garrison, S. M., Gill, T., Hahn, A. C., Jaeger, B., Kačmár, P., Kaminski, G., Kanske, P., Kekecs, Z., Kline, M., Koehn, M. A., Kujur, P., Levitan, C. A., Miller, J. K., Okan, C., Olsen, J., Oviedo-Trespalacios, O., Özdoğru, A. A., Pande, B., Parganiha, A., Parveen, N., Pfuhl, G., Pradhan, S., Ropovik, I., Rule, N. O., Saunders, B., Schei, V., Schmidt, K., Singh, M. M., Sirota, M., Steltenpohl, C. N., Stieger, S., Storage, D., Sullivan, G. B., Szabelska, A., Tamnes, C. K., Vadillo, M. A., Valentova, J. V., Vanpaemel, W., Varella, M. A. C., Vergauwe, E., Verschoor, M., Vianello, M., Voracek, M., Williams, G. P., Wilson, J. P., Zickfeld, J. H., Arnal, J. D., Aydin, B., Chen, S.-C., DeBruine, L. M., Fernandez, A. M., Horstmann, K. T., Isager, P. M., Jones, B., Kapucu, A., Lin, H., Mensink, M. C., Navarrete, G., Silan, M. A., and Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science*, 1(4):501–515.

Mulgan, R. (2000). 'Accountability': An Ever-Expanding Concept? *Public Administration*, 78(3):555–573.

Nisbett, R. E., Zukier, H., and Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, 13(2):248–277.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer,

A. E., and Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*, 23(10):815–818.

Novak, T. P. and Hoffman, D. L. (2009). The Fit of Thinking Style and Situation: New Measures of Situation-Specific Experiential and Rational Cognition. *Journal of Consumer Research*, 36(1):56–72.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716–aac4716.

Orquin, J. L. and Mueller Loose, S. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, 144(1):190–206.

Pachur, T. and Marinello, G. (2013). Expert intuitions: How to model the decision strategies of airport customs officers? *Acta Psychologica*, 144(1):97–103.

Patil, S. V., Tetlock, P. E., and Mellers, B. A. (2017). Accountability Systems and Group Norms: Balancing the Risks of Mindless Conformity and Reckless Deviation: Process versus Outcome Accountability. *Journal of Behavioral Decision Making*, 30(2):282–303.

Patil, S. V., Vieider, F., and Tetlock, P. E. (2014). Process Versus Outcome Accountability. In Bovens, M., Goodin, R. E., and Schillemans, T., editors, *The Oxford Handbook of Public Accountability*. Oxford University Press.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge University Press.

Peecher, M. E., Solomon, I., and Trotman, K. T. (2013). An accountability framework for financial statement auditors and related research questions. *Accounting, Organizations and Society*, 38(8):596–620.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2):8–13.

Pennington, J. and Schlenker, B. R. (1999). Accountability for Consequential Decisions: Justifying Ethical Judgments to Audiences. *Personality and Social Psychology Bulletin*, 25(9):1067–1081.

Pitesa, M. and Thau, S. (2013). Masters of the universe: How power and accountability influence self-serving decisions under moral hazard. *Journal of Applied Psychology*, 98(3):550–558.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Working Papers*, page 8.

Popper, K. (2005). *The Logic of Scientific Discovery*. Routledge.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, 25:111–163.

Rausch, A. and Brauneis, A. (2015). The effect of accountability on management accountants' selection of information. *Review of Managerial Science*, 9(3):487–521.

Reverberi, C., Toraldo, A., D'Agostini, S., and Skrap, M. (2005). Better without (lateral) frontal cortex? Insight problems solved by frontal patients. *Brain*, 128(12):2882–2890.

Rieskamp, J. (2006). Perspectives of probabilistic inferences: Reinforcement learning and an adaptive network compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(6):1355–1370.

Rieskamp, J. and Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, 127(2):258–276.

Rieskamp, J. and Otto, P. E. (2006). SSL: A Theory of How People Learn to Select Strategies. *Journal of Experimental Psychology: General*, 135(2):207–236.

Rodgers, W. (1999). The influences of conflicting information on novices and loan officers' actions. *Journal of Economic Psychology*, 20(2):123–145.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641.

Runco, M. A. and Jaeger, G. J. (2012). The Standard Definition of Creativity. *Creativity Research Journal*, 24(1):92–96.

Russo, J. E., Meloy, M. G., and Wilks, T. J. (2000). Predecisional Distortion of Information by Auditors and Salespersons. *Management Science*, 46(1):13–27.

Schillemans, T. (2016). Calibrating Public Sector Accountability: Translating experimental findings to public sector accountability. *Public Management Review*, 18(9):1400–1420.

Scholten, L., van Knippenberg, D., Nijstad, B. A., and De Dreu, C. K. (2007). Motivated information processing and group decision-making: Effects of process accountability on information processing and decision quality. *Journal of Experimental Social Psychology*, 43(4):539–552.

Schooler, J. W., Ohlsson, S., and Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of experimental psychology: General*, 122(2):166.

Schooler, L. J. and Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, 112(3):610–628.

Schulte-Mecklenbeck, M., Johnson, J. G., Böckenholt, U., Goldstein, D. G., Russo, J. E., Sullivan, N. J., and Willemsen, M. C. (2017). Process-Tracing Methods in Decision Making: On Growing Up in the 70s. *Current Directions in Psychological Science*, 26(5):442–450.

Schulz-Hardt, S., Rollwage, J., Wanzel, S. K., Frisch, J. U., and Häusser, J. A. (2020). Effects of process and outcome accountability on escalating commitment: A two-study replication. *Journal of Experimental Psychology: Applied*.

Sedikides, C., Herbst, K. C., Hardin, D. P., and Dardis, G. J. (2002). Accountability as a deterrent to self-enhancement: The search for mechanisms. *Journal of Personality and Social Psychology*, 83(3):592–605.

Shah, A. K. and Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, 134(2):207–222.

Siegel-Jacobs, K. and Yates, J. F. (1996). Effects of procedural and outcome accountability on judgment quality. *Organizational Behavior and Human Decision Processes*, 65(1):1–17.

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

Simonson, I. and Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes*, 51(3):416–446.

Simonson, I. and Staw, B. M. (1992). Deescalation strategies: A comparison of techniques for reducing commitment to losing courses of action. *Journal of Applied Psychology*, 77(4):419–426.

Şimşek, Ö. (2013). Linear decision rule as aspiration for simple decision heuristics. In *Advances in Neural Information Processing Systems*, pages 2904–2912.

Slaughter, J. E., Bagger, J., and Li, A. (2006). Context effects on group-based employee selection decisions. *Organizational Behavior and Human Decision Processes*, 100(1):47–59.

Smaldino, P. E. and McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9):160384.

Söllner, A., Bröder, A., and Hilbig, B. E. (2013). Deliberation versus automaticity in decision making: Which presentation format features facilitate automatic decision making? *Judgment and Decision Making*, 8(3):278.

Spellman, B. A. (2015). A Short (Personal) Future History of Revolution 2.0. *Perspectives on Psychological Science*, 10(6):886–899.

Stan development team (2017). Stan Modeling Language: User's Guide and Reference Manual, Version 2.17.0.

Starcke, K. and Brand, M. (2012). Decision making under stress: A selective review. *Neuroscience & Biobehavioral Reviews*, 36(4):1228–1248.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, 11(5):702–712.

Stein, M. I. (1953). Creativity and Culture. *The Journal of Psychology*, 36(2):311–322.

Studdert, D. M. (2005). Defensive Medicine Among High-Risk Specialist Physicians in a Volatile Malpractice Environment. *JAMA*, 293(21):2609.

Taylor, D. W., Berry, P. C., and Block, C. H. (1958). Does Group Participation When Using Brainstorming Facilitate or Inhibit Creative Thinking? *Administrative Science Quarterly*, 3(1):23.

Tetlock, P. and Kim, J. I. (1987). Accountability and Judgment Processes in a Personality Prediction Task | Ovid. *Journal of Personality and Social Psychology*, 52(4):700–709.

Tetlock, P. E. (1983a). Accountability and complexity of thought. *Journal of Personality and Social Psychology*, 45(1):74–83.

Tetlock, P. E. (1983b). Accountability and the Perseverance of First Impressions. *Social Psychology Quarterly*, 46(4):285–292.

Tetlock, P. E. (1985). Accountability: The neglected social context of judgment and choice. *Research in organizational behavior*, 7(1):297–332.

Tetlock, P. E. (1992). The Impact of Accountability on Judgment and Choice: Toward A Social Contingency Model. In *Advances in Experimental Social Psychology*, volume 25, pages 331–376. Elsevier.

Tetlock, P. E. and Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of personality and social psychology*, 57(3):388.

Tetlock, P. E. and Lerner, J. S. (1999). The Social Contingency Model: Identifying Empirical and Normative Boundary Conditions on the Error-and-Bias Portrait of Human Nature. In *Dual Process Theories in Social Psychology*, page 35. Guilford Press, New York.

Tetlock, P. E., Lerner, J. S., and Boettger, R. (1996). The dilution effect: Judgmental bias, conversational convention, or a bit of both? *European Journal of Social Psychology*, 26(6):915–934.

Tetlock, P. E., Skitka, L., and Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. *Journal of personality and social psychology*, 57(4):632.

Tetlock, P. E., Vieider, F. M., Patil, S. V., and Grant, A. M. (2013). Accountability and ideology: When left looks right and right looks left. *Organizational Behavior and Human Decision Processes*, 122(1):22–35.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving Decisions about Health, Wealth and Happiness*. Penguin Books, London.

Todd, P. M. and Brighton, H. (2016). Building the Theory of Ecological Rationality. *Minds and Machines*, 26(1-2):9–30.

Todd, P. M. and Gigerenzer, G. (2007). Environments That Make Us Smart: Ecological Rationality. *Current Directions in Psychological Science*, 16(3):167–171.

Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131.

Vandegrift, D. and Brown, P. (2003). Task difficulty, incentive effects, and the selection of high-variance strategies: An experimental examination of tournament behavior. *Labour Economics*, 10(4):481–497.

Verwaeren, B. (2019). Controlling the What or the How: The Effect of Accountability Focus on Individual Exploration.

Vieider, F. M. (2009). The effect of accountability on loss aversion. *Acta Psychologica*, 132(1):96–101.

Vieider, F. M. (2011). Separating real incentives and accountability. *Experimental Economics*, 14(4):507–518.

Vivian Chen, C.-H., Yuan, M.-L., Cheng, J.-W., and Seifert, R. (2016). Linking transformational leadership and core self-evaluation to job performance: The mediating role of felt accountability. *The North American Journal of Economics and Finance*, 35:234–246.

Walsh, M. M. and Gluck, K. A. (2016). Verbalization of Decision Strategies in Multiple-Cue Probabilistic Inference: Verbalization of Decision Strategies. *Journal of Behavioral Decision Making*, 29(1):78–91.

Wegwarth, O., Gaissmaier, W., and Gigerenzer, G. (2009). Smart strategies for doctors and doctors-in-training: Heuristics in medicine. *Medical Education*, 43(8):721–728.

Wiley, J. and Jarosz, A. F. (2012). Working Memory Capacity, Attentional Focus, and Problem Solving. *Current Directions in Psychological Science*, 21(4):258–262.

Woike, J. K., Hoffrage, U., and Martignon, L. (2017). Integrating and testing natural frequencies, naïve Bayes, and fast-and-frugal trees. *Decision*, 4(4):234–260.

Wübben, M. and Wangenheim, F. (2008). Instant Customer Base Analysis: Managerial Heuristics Often "Get It Right". *Journal of Marketing*, 72(3):13.

Zedelius, C. M. and Schooler, J. W. (2015). Mind wandering "Ahas" versus mindful reasoning: Alternative routes to creative solutions. *Frontiers in Psychology*, 6.

Zwaan, R. A., Etz, A., Lucas, R. E., and Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41.

Summary

Accurate judgment, adaptive decision making, and ability to find insightful solutions to challenging problems are some of the key qualities organizations aspire to instill in their members. In the service of this goal, designers of management systems draw on a variety of approaches, ranging from formal contracts to subtle "nudges." At the core of many of these approaches is the idea of *accountability*, the use of external scrutiny to bring about desired behavior and outcomes. How should accountability be implemented to elicit optimal performance from individuals? Does scrutinizing the *process* the accountable individual is using tend to yield more accurate judgments and decisions than focusing strictly on the *outcome*? Does the answer depend on the nature of the decision environment? Does the answer hold in the domain of problems that call for insightful solutions? What are the cognitive and affective mechanisms that mediate the differences in performance between *process accountability* and *outcome accountability*? Are these effects robust or do they depend on currently insufficiently understood moderating factors? This dissertation reports the results of a series of behavioral experiments designed to shed light on these and related questions.

Chapters 2, 3, and 4 investigate the differences in performance between process and outcome accountability. Chapter 2 explores multiple-cue judgment, seeking to replicate and extend the finding that process accountability results in superior performance compared to outcome accountability and that this effect is mediated by the more deliberate and systematic approach of process-accountable individuals. The two chapters that follow seek to identify boundary conditions for the superiority of process accountability by investigating insight problem solving (Chapter 3) and adaptive decision making (Chapter 4). Apart from investigating performance differences, these studies draw on self-report and process-tracing measures to shed light on the underlying cognitive and affective processes. Overall, the results of these studies paint a more favorable picture of outcome accountability compared to previous research, highlighting boundary conditions for the superiority of process accountability and suggesting that some of the reported results (concerning both performance and cognitive processing effects) may be more variable than previously thought.

Finally, the research reported in Chapter 5 combines cognitive modeling and analysis of eye-tracking data to investigate the decision-making strategies of professional loan officers who operate in an environment in which the requirement to justify one's decisions looms large. The results of latent-mixture modeling suggest that loan officers tend to rely on informative-intensive, compensatory decision-making strategies, and the analysis of information search patterns afforded by eye-tracking data provides convergent evidence.

The work reported in this dissertation informs the debates about how accountability should be structured to facilitate performance. But it also suggests that some changes to research practice

would be beneficial. An important takeaway is that we should take seriously the possibility that the reported findings are highly sensitive to variation in accountability implementation and context. Independent replication attempts should be more common than they currently are. Other developments that would facilitate cumulative progress include open science practices, reuse of research materials, and larger-scale collaborative research. These practices would speed up progress and make it more likely that academic work on accountability will, in time, produce precise, robust, and genuinely useful guidelines for management system designers implementing accountability in their organizations.

Samenvatting

Nauwkeurig oordelen, adaptieve besluitvorming en het vermogen om inzichtrijke oplossingen voor moeilijke problemen te vinden behoren tot de cruciale kwaliteiten die organisaties in hun mensen proberen te bevorderen. Om dit doel te bereiken maken ontwerpers van managementsystemen gebruik van verschillende benaderingen, van formele contracten tot subtiele "nudges." Centraal in veel van deze benaderingen staat het idee van *accountability*, het gebruik van externe onderzoeking om gewenst gedrag en uitkomsten te bewerkstelligen. Hoe moet accountability geïmplementeerd worden om optimale prestaties uit personen te halen? Levert het examineren van het *proces* dat de persoon gebruikt die accountable is meer nauwkeurige oordelen en beslissingen op dan als men focust op de *uitkomst*? Hangt het antwoord af van de aard van de beslissingsomgeving? Houdt dat antwoord ook stand in het domein van problemen waar inzicht-oplossingen nodig zijn? Wat zijn de cognitieve en affectieve mechanismen die de verschillen in prestaties tussen proces-accountability en uitkomst-accountability mediëren? Zijn deze effect robuust of hangen deze af van momenteel onvoldoende bekende modererende factoren? Dit proefschrift biedt de resultaten van een serie gedragsexperimenten die ontworpen zijn om deze en aanverwante vragen te beantwoorden.

Hoofdstukken 2, 3 en 4 onderzoeken de verschillen in prestaties tussen proces- en uitkomst-accountability. Hoofdstuk 2 verkent de situatie van oordelen met meerdere aanwijzingen in een poging de bevinding te repliceren en uit te breiden dat proces-accountability hogere prestaties oplevert dan uitkomst-accountability en dat dit effect gemedieerd wordt door de meer intentionele en systematische aanpak van proces-accountable personen. De twee daaropvolgende hoofdstukken pogen de voorwaarden voor de relatief positieve effecten van proces-accountability in kaart te brengen door inzichtrijke probleemoplossingen (Hoofdstuk 3) en adaptieve besluitvorming (Hoofdstuk 4) te onderzoeken. Naast het onderzoeken van verschillen in prestaties gebruiken deze studies zelf-gerapporteerde en proces-gerelateerde metingen om de onderliggende cognitieve en affectieve processen te belichten. Samengenomen bieden de resultaten van deze onderzoeken een positiever beeld van uitkomst-accountability dan eerder onderzoek, door de voorwaarden voor de betere resultaten van proces-accountability te belichten en door de mogelijkheid te overwegen dat sommige resultaten (omtrent zowel prestaties als cognitieve effecten) meer variëren dan eerder werd aangenomen.

Tot slot combineert het onderzoek in Hoofdstuk 5 cognitieve modellen en een analyse van eye-trackingdata om de besluitvormingsstrategieën te onderzoeken van professionele kredietverstrekkers die werken in een omgeving waar de eis om hun beslissingen te verantwoorden nadrukkelijk aanwezig is. De resultaten van latent-mixture-modellen suggereren dat kredietverstrekkers vertrouwen op informatie-intense, compenserende besluitvormingsstrategieën, en de analyse van patronen in het doorzoeken van informatie die door eye-tracking data

mogelijk gemaakt wordt, biedt convergerend bewijs.

Het onderzoek dat in deze dissertatie gepresenteerd wordt, informeert het debat over hoe accountability gestructureerd moet worden om goede prestaties te bevorderen. Maar het suggereert ook dat enkele veranderingen in de manier van onderzoek doen nuttig kunnen zijn. Een belangrijke conclusie is dat we de mogelijkheid serieus moeten nemen dat gepubliceerde vondsten erg gevoelig zijn voor verschillen in de implementatie van accountability en in de context. Onafhankelijke pogingen tot replicatie zouden meer gebruikelijk moeten zijn dan ze nu zijn. Andere ontwikkelingen die cumulatieve vooruitgang kunnen bevorderen zijn *open science*-benaderingen, het hergebruiken van onderzoeksmaterialen en grootschaliger samenwerking in onderzoek. Deze methoden zouden de vooruitgang versnellen en het waarschijnlijker maken dat academisch werk over accountability met de tijd precieze, robuuste en werkelijk nuttige richtlijnen opleveren voor de ontwerpers van managementsystemen die accountability in hun organisaties implementeren.

About the author

Aljaž Sluga was born in Postojna, Slovenia in 1989. He received his degree in psychology from the University of Ljubljana in 2014. After graduation and a temporary gig as a market research analyst, he worked as a researcher at the Mind & Brain Lab at the University of Ljubljana.

In 2015, he started his PhD at the Rotterdam School of Management and the Erasmus Research Institute of Management, investigating accountability under the supervision of Prof. dr. Frank G. H. Hartmann and Dr. Maarten A. S. Boksem. During the PhD, he attended and presented at academic conferences and conducted workshops in academic and business settings.

In the academic year 2021/2021, he has a visiting position at the Rotterdam School of Management. In 2021, he has started working as a data scientist at Studio Moderna.

Portfolio

Papers in preparation

- Process and outcome accountability in multiple-cue judgment: two replication studies
- Decision making strategies of loan officers: modeling and eye-tracking evidence

Teaching

Thesis supervision

- Bachelor thesis supervision (2020)
- Master thesis supervision (2016 - 2020)

Courses

- Neuroaccounting (2016-2017)
- Neuroaccounting and finance (2018)
- Accounting Analytics (2019 - 2020)

Coursework

Regular ERIM courses

- Behavioral decision theory
- Programming
- English
- Presentation skills
- Statistical methods
- Developing theory and theoretical contributions
- Advanced statistical methods
- Philosophy of science
- Scientific integrity
- Publishing strategy

ERIM summer school courses

- My first Bayes
- Testing and interpreting moderation and mediation with SPSS
- Methods stumblers

The rest

- Gathering data to test accounting theories (Amsterdam, 2015)
- Lund eye-tracking academy (Lund, 2017)
- Applying multinomial processing tree models in experimental research (Rotterdam, 2017)
- Bayesian modeling for cognitive science (Amsterdam, 2017)
- Agent-based modeling (Eindhoven, 2017)

ERIM PhD Series

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://repub.eur.nl/pub. ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam (EUR).

Dissertations in the last four years

Ahmadi, S., *A motivational perspective to decision-making and behavior in organizations*, Promotors: Prof. J.J.P. Jansen & Dr T.J.M. Mom, EPS-2019-477-S&E, https://repub.eur.nl/pub/116727

Akemu, O., *Corporate Responses to Social Issues: Essays in Social Entrepreneurship and Corporate Social Responsibility*, Promotors: Prof. G.M. Whiteman & Dr S.P. Kennedy, EPS-2017-392-ORG, https://repub.eur.nl/pub/95768

Albuquerque de Sousa, J.A., *International stock markets: Essays on the determinants and consequences of financial market development*, Promotors: Prof. M.A. van Dijk & Prof. P.A.G. van Bergeijk, EPS-2019-465-F&A, https://repub.eur.nl/pub/115988

Alserda, G.A.G., *Choices in Pension Management*, Promotors: Prof. S.G. van der Lecq & Dr O.W. Steenbeek, EPS-2017-432-F&A, https://repub.eur.nl/pub/103496

Arampatzi, E., *Subjective Well-Being in Times of Crises: Evidence on the Wider Impact of Economic Crises and Turmoil on Subjective Well-Being*, Promotors: Prof. H.R. Commandeur, Prof. F. van Oort & Dr. M.J. Burger, EPS-2018-459-S&E, https://repub.eur.nl/pub/111830

Arslan, A.M., *Operational Strategies for On-demand Delivery Services*, Promotors: Prof. R.A. Zuidwijk & Dr N.A. H. Agatz, EPS-2019-481-LIS, https://repub.eur.nl/pub/126463

Avci, E., *Surveillance of Complex Auction Markets: a Market Policy Analytics Approach*, Promotors: Prof. W. Ketter, Prof. H.W.G.M. van Heck & Prof. D.W. Bunn, EPS-2018-426-LIS, https://repub.eur.nl/pub/106286

Balen, T.H. van, *Challenges of Early Stage Entrepreneurs: the Roles of Vision Communication and Team Membership Change*, Promotors: Prof. J.C.M. van den Ende & Dr M. Tarakci,

EPS-2019-468-LIS, https://repub.eur.nl/pub/115654

Bansraj, S.C., *The Principles of Private Equity: Ownership and Acquisitions*, Promotors: Prof. J.T.J Smit & Dr V. Volosovych, EPS-2020-507-F&A, https://repub.eur.nl/pub/132329

Bavato, D., *With New Eyes: The recognition of novelty and novel ideas*, Promotors: Prof. D.A. Stam & Dr. S. Tasselli, EPS-2020-500-LIS, https://repub.eur.nl/pub/134264

Bernoster, I., *Essays at the Intersection of Psychology, Biology, and Entrepreneurship*, Promotors: Prof. A.R. Thurik, Prof. I.H.A. Franken & Prof. P.J.F Groenen, EPS-2018-463-S&E, https://repub.eur.nl/pub/113907

Blagoeva, R.R., *The Hard Power Of Soft Power: A behavioral strategy perspective on how power, reputation, and status affect firms*, Promotors: Prof. J.J.P. Jansen & Prof. T.J.M. Mom, EPS-2020-495-S&E, https://repub.eur.nl/pub/127681

Bouman, P., *Passengers, Crowding and Complexity: Models for Passenger Oriented Public Transport*, Prof. L.G. Kroon, Prof. A. Schöbel & Prof. P.H.M. Vervest, EPS-2017-420-LIS, https://repub.eur.nl/pub/100767

Breugem, T., *Crew Planning at Netherlands Railways: Improving Fairness, Attractiveness, and Efficiency*, Promotors: Prof. D. Huisman & Dr T.A.B. Dollevoet, EPS-2020-494-LIS, https://repub.eur.nl/pub/124016

Bunderen, L. van, *Tug-of-War: Why and when teams get embroiled in power struggles*, Promotors: Prof. D.L. van Knippenberg & Dr. L. Greer, EPS-2018-446-ORG, https://repub.eur.nl/pub/105346

Burg, G.J.J. van den, *Algorithms for Multiclass Classification and Regularized Regression*, Promotors: Prof. P.J.F. Groenen & Dr. A. Alfons, EPS-2018-442-MKT, https://repub.eur.nl/pub/103929

Chammas, G., *Portfolio concentration*, Promotor: Prof. J. Spronk, EPS-2017-410-F&E, https://repub.eur.nl/pub/94975

Chan, H.Y., *Decoding the consumer's brain: Neural representations of consumer experience*, Promotors: Prof. A. Smidts & Dr M.A.S. Boksem, EPS-2019-493-MKT, https://repub.eur.nl/pub/124931

Couwenberg, L., *Context dependent valuation: A neuroscientific perspective on consumer decision-making*, Promotors: Prof. A. Smit, Prof. A.G. Sanfrey & Dr M.A.S. Boksem, EPS-2020-505-MKT, https://repub.eur.nl/pub/129601

Dalmeijer, K., *Time Window Assignment in Distribution Networks*, Promotors: Prof A.P.M. Wagelmans & Dr R. Spliet, EPS-2019-486-LIS, https://repub.eur.nl/pub/120773

Dennerlein, T. *Empowering Leadership and Employees' Achievement Motivations: the Role of Self-Efficacy and Goal Orientations in the Empowering Leadership Process*, Promotors: Prof. D.L. van Knippenberg & Dr J. Dietz, EPS-2017-414-ORG, https://repub.eur.nl/pub/98438

Dolgova, E., *On Getting Along and Getting Ahead: How Personality Affects Social Network Dynamics*, Promotors: Prof. P.P.M.A.R Heugens & Prof. M.C. Schippers, EPS-2019-455-S&E,

https://repub.eur.nl/pub/119150

Duijzer, L.E., *Mathematical Optimization in Vaccine Allocation*, Promotors: Prof. R. Dekker & Dr W.L. van Jaarsveld, EPS-2017-430-LIS, https://repub.eur.nl/pub/101487

Fasaei, H., *Changing the Narrative: The Behavioral Effects of Social Evaluations on the Decision Making of Organizations*, Promotors: Prof. J.J.P. Jansen, Prof. T.J.M. Mom & Dr. M.P. Tempelaar, EPS-2020-492-S&E, https://repub.eur.nl/pub/129598

Eijlers, E., *Emotional Experience and Advertising Effectiveness: on the use of EEG in marketing*, Prof. A. Smidts & Prof. M.A.S. Boksem, EPS-2019-487-MKT, https://repub.eur.nl/pub/124053

El Nayal, O.S.A.N., *Firms and the State: An Examination of Corporate Political Activity and the Business-Government Interface*, Promotor: Prof. J. van Oosterhout & Dr. M. van Essen, EPS-2018-469-S&E, https://repub.eur.nl/pub/114683

Feng, Y., *The Effectiveness of Corporate Governance Mechanisms and Leadership Structure: Impacts on strategic change and firm performance*, Promotors: Prof. F.A.J. van den Bosch, Prof. H.W. Volberda & Dr J.S. Sidhu, EPS-2017-389-S&E, https://repub.eur.nl/pub/98470

Frick, T.W., *The Implications of Advertising Personalization for Firms, Consumer, and Ad Platfroms*, Promotors: Prof. T. Li & Prof. H.W.G.M. van Heck, EPS-2018-452-LIS, https://repub.eur.nl/pub/110314

Fytraki, A.T., *Behavioral Effects in Consumer Evaluations of Recommendation Systems*, Promotors: Prof. B.G.C. Dellaert & Prof. T. Li, EPS-2018-427-MKT, https://repub.eur.nl/pub/110457

Gai, J., *Contextualized Consumers: Theories and Evidence on Consumer Ethics, Product Recommendations, and Self-Control*, Promotors: Prof. S. Puntoni & Prof. S.T.L. Sweldens, EPS-2020-498-MKT, https://repub.eur.nl/pub/127680

Ghazizadeh, P. *Empirical Studies on the Role of Financial Information in Asset and Capital Markets*, Promotors: Prof. A. de Jong & Prof. E. Peek, EPS-2019-470-F&A, https://repub.eur.nl/pub/114023

Giurge, L., *A Test of Time; A temporal and dynamic approach to power and ethics*, Promotors: Prof. M.H. van Dijke & Prof. D. De Cremer, EPS-2017-412-ORG, https://repub.eur.nl/pub/98451

Gobena, L., *Towards Integrating Antecedents of Voluntary Tax Compliance*, Promotors: Prof. M.H. van Dijke & Dr P. Verboon, EPS-2017-436-ORG, https://repub.eur.nl/pub/103276

Groot, W.A., *Assessing Asset Pricing Anomalies*, Promotors: Prof. M.J.C.M. Verbeek & Prof. J.H. van Binsbergen, EPS-2017-437-F&A, https://repub.eur.nl/pub/103490

Hanselaar, R.M., *Raising Capital: On pricing, liquidity and incentives*, Promotors: Prof. M.A. van Dijk & Prof. P.G.J. Roosenboom, EPS-2018-429-F&A, https://repub.eur.nl/pub/113274

Harms, J. A., *Essays on the Behavioral Economics of Social Preferences and Bounded Rationality*, Prof. H.R. Commandeur & Dr K.E.H. Maas, EPS-2018-457-S&E, https://repub.eur.nl/pub/108831

Hendriks, G., *Multinational Enterprises and Limits to International Growth: Links between Domestic and Foreign Activities in a Firm's Portfolio*, Promotors: Prof. P.P.M.A.R. Heugens & Dr. A.H.L Slangen, EPS-2019-464-S&E, https://repub.eur.nl/pub/114981

Hengelaar, G.A., *The Proactive Incumbent: Holy grail or hidden gem? Investigating whether the Dutch electricity sector can overcome the incumbent's curse and lead the sustainability transition*, Promotors: Prof. R.J. M. van Tulder & Dr K. Dittrich, EPS-2018-438-ORG, https://repub.eur.nl/pub/102953

Jacobs, B.J.D., *Marketing Analytics for High-Dimensional Assortments*, Promotors: Prof. A.C.D. Donkers & Prof. D. Fok, EPS-2017-445-MKT, https://repub.eur.nl/pub/103497

Jia, F., *The Value of Happiness in Entrepreneurship*, Promotors: Prof. D.L. van Knippenberg & Dr Y. Zhang, EPS-2019-479-ORG, https://repub.eur.nl/pub/115990

Kahlen, M. T., *Virtual Power Plants of Electric Vehicles in Sustainable Smart Electricity Markets*, Promotors: Prof. W. Ketter & Prof. A. Gupta, EPS-2017-431-LIS, https://repub.eur.nl/pub/100844

Kampen, S. van, *The Cross-sectional and Time-series Dynamics of Corporate Finance: Empirical evidence from financially constrained firms*, Promotors: Prof. L. Norden & Prof. P.G.J. Roosenboom, EPS-2018-440-F&A, https://repub.eur.nl/pub/105245

Karali, E., *Investigating Routines and Dynamic Capabilities for Change and Innovation*, Promotors: Prof. H.W. Volberda, Prof. H.R. Commandeur & Dr J.S. Sidhu, EPS-2018-454-S&E, https://repub.eur.nl/pub/106274

Keko. E, *Essays on Innovation Generation in Incumbent Firms*, Promotors: Prof. S. Stremersch & Dr N.M.A. Camacho, EPS-2017-419-MKT, https://repub.eur.nl/pub/100841

Kerkkamp, R.B.O., *Optimisation Models for Supply Chain Coordination under Information Asymmetry*, Promotors: Prof. A.P.M. Wagelmans & Dr. W. van den Heuvel, EPS-2018-462-LIS, https://repub.eur.nl/pub/109770

Khattab, J., *Make Minorities Great Again: a contribution to workplace equity by identifying and addressing constraints and privileges*, Promotors: Prof. D.L. van Knippenberg & Dr A. Nederveen Pieterse, EPS-2017-421-ORG, https://repub.eur.nl/pub/99311

Kim, T. Y., *Data-driven Warehouse Management in Global Supply Chains*, Promotors: Prof. R. Dekker & Dr C. Heij, EPS-2018-449-LIS, https://repub.eur.nl/pub/109103

Klitsie, E.J., *Strategic Renewal in Institutional Contexts: The paradox of embedded agency*, Promotors: Prof. H.W. Volberda & Dr. S. Ansari, EPS-2018-444-S&E, https://repub.eur.nl/pub/106275

Koolen, D., *Market Risks and Strategies in Power Systems Integrating Renewable Energy*, Promotors: Prof. W. Ketter & Prof. R. Huisman, EPS-2019-467-LIS, https://repub.eur.nl/pub/115655

Kong, L. *Essays on Financial Coordination*, Promotors: Prof. M.J.C.M. Verbeek, Dr. D.G.J. Bongaerts & Dr. M.A. van Achter. EPS-2019-433-F&A, https://repub.eur.nl/pub/114516

Kyosev, G.S., *Essays on Factor Investing*, Promotors: Prof. M.J.C.M. Verbeek & Dr J.J. Huij, EPS-2019-474-F&A, https://repub.eur.nl/pub/116463

Lamballais Tessensohn, T., *Optimizing the Performance of Robotic Mobile Fulfillment Systems*, Promotors: Prof. M.B.M de Koster, Prof. R. Dekker & Dr D. Roy, EPS-2019-411-LIS, https://repub.eur.nl/pub/116477

Leung, W.L., *How Technology Shapes Consumption: Implications for Identity and Judgement*, Promotors: Prof. S. Puntoni & Dr G Paolacci, EPS-2019-485-MKT, https://repub.eur.nl/pub/117432

Li, X. *Dynamic Decision Making under Supply Chain Competition*, Promotors: Prof. M.B.M de Koster, Prof. R. Dekker & Prof. R. Zuidwijk, EPS-2018-466-LIS, https://repub.eur.nl/pub/114028

Liu, N., *Behavioral Biases in Interpersonal Contexts*, Supervisors: Prof. A. Baillon & Prof. H. Bleichrodt, EPS-2017-408-MKT, https://repub.eur.nl/pub/95487

Maas, A.J.J., *Organizations and their external context: Impressions across time and space*, Promotors: Prof. P.P.M.A.R Heugens & Prof. T.H. Reus, EPS-2019-478-S&E, https://repub.eur.nl/pub/116480

Maira, E., *Consumers and Producers*, Promotors: Prof. S. Puntoni & Prof. C. Fuchs, EPS-2018-439-MKT, https://repub.eur.nl/pub/104387

Manouchehrabadi, B., *Information, Communication and Organizational Behavior*, Promotors: Prof. G.W.J. Hendrikse & Dr O.H. Swank, EPS-2020-502-ORG, https://repub.eur.nl/pub/132185

Matawlie, N., *Through Mind and Behaviour to Financial Decisions*, Promotors: Prof. J.T.J. Smit & Prof. P. Verwijmeren, EPS-2020-501-F&A, https://repub.eur.nl/pub/134265

Mirzaei, M., *Advanced Storage and Retrieval Policies in Automated Warehouses*, Promotors: Prof. M.B.M. de Koster & Dr N. Zaerpour, EPS-2020-490-LIS, https://repub.eur.nl/pub/125975

Nair, K.P., *Strengthening Corporate Leadership Research: The relevance of biological explanations*, Promotors: Prof. J. van Oosterhout & Prof. P.P.M.A.R Heugens, EPS-2019-480-S&E, https://repub.eur.nl/pub/120023

Nullmeier, F.M.E., *Effective contracting of uncertain performance outcomes: Allocating responsibility for performance outcomes to align goals across supply chain actors*, Promotors: Prof. J.Y.F.Wynstra & Prof. E.M. van Raaij, EPS-2019-484-LIS, https://repub.eur.nl/pub/118723

Okbay, A., *Essays on Genetics and the Social Sciences*, Promotors: Prof. A.R. Thurik, Prof. Ph.D. Koellinger & Prof. P.J.F. Groenen, EPS-2017-413-S&E, https://repub.eur.nl/pub/95489

Peng, X., *Innovation, Member Sorting, and Evaluation of Agricultural Cooperatives*, Promotor: Prof. G.W.J. Hendriks, EPS-2017-409-ORG, https://repub.eur.nl/pub/94976

Petruchenya, A., *Essays on Cooperatives: Emergence, Retained Earnings, and Market Shares*, Promotors: Prof. G.W.J. Hendriks & Dr Y. Zhang, EPS-2018-447-ORG, https://repub.eur.nl/pub/105243

Plessis, C. du, *Influencers: The Role of Social Influence in Marketing*, Promotors: Prof. S. Puntoni & Prof. S.T.L.R. Sweldens, EPS-2017-425-MKT, https://repub.eur.nl/pub/103265

Pocock, M., *Status Inequalities in Business Exchange Relations in Luxury Markets*, Promotors: Prof. C.B.M. van Riel & Dr G.A.J.M. Berens, EPS-2017-346-ORG, https://repub.eur.nl/pub/98647

Polinder, G.J., *New Models and Applications for Railway Timetabling*, Prof. D. Huisman & Dr. M.E. Schmidt, EPS-2020-514-LIS, https://repub.eur.nl/pub/134600

Pozharliev, R., *Social Neuromarketing: The role of social context in measuring advertising effectiveness*, Promotors: Prof. W.J.M.I. Verbeke & Prof. J.W. van Strien, EPS-2017-402-MKT, https://repub.eur.nl/pub/95528

Qian, Z., *Time-Varying Integration and Portfolio Choices in the European Capital Markets*, Promotors: Prof. W.F.C. Verschoor, Prof. R.C.J. Zwinkels & Prof. M.A. Pieterse-Bloem, EPS-2020-488-F&A, https://repub.eur.nl/pub/124984

Reh, S.G., *A Temporal Perspective on Social Comparisons in Organizations*, Promotors: Prof. S.R. Giessner, Prof. N. van Quaquebeke & Dr. C. Troster, EPS-2018-471-ORG, https://repub.eur.nl/pub/114522

Riessen, B. van, *Optimal Transportation Plans and Portfolios for Synchromodal Container Networks*, Promotors: Prof. R. Dekker & Prof. R.R. Negenborn, EPS-2018-448-LIS, https://repub.eur.nl/pub/105248

Romochkina, I.V., *When Interests Collide: Understanding and modeling interests alignment using fair pricing in the context of interorganizational information systems*, Promotors: Prof. R.A. Zuidwijk & Prof. P.J. van Baalen, EPS-2020-451-LIS, https://repub.eur.nl/pub/127244

Schie, R. J. G. van, *Planning for Retirement: Save More or Retire Later?* Promotors: Prof. B. G. C. Dellaert & Prof. A.C.D. Donkers, EOS-2017-415-MKT, https://repub.eur.nl/pub/100846

Schneidmüller, T., *Engaging with Emerging Technologies: Socio-cognitive foundations of incumbent response*, Promotors: Prof. H. Volberda & Dr S.M. Ansari, EPS-2020-509-S&E, https://repub.eur.nl/pub/131124

Schouten, K.I.M. *Semantics-driven Aspect-based Sentiment Analysis*, Promotors: Prof. F.M.G. de Jong, Prof. R. Dekker & Dr. F. Frasincar, EPS-2018-453-LIS, https://repub.eur.nl/pub/112161

Sihag, V., *The Effectiveness of Organizational Controls: A meta-analytic review and an investigation in NPD outsourcing*, Promotors: Prof. J.C.M. van den Ende & Dr S.A. Rijsdijk, EPS-2019-476-LIS, https://repub.eur.nl/pub/115931

Smolka, K.M., *Essays on Entrepreneurial Cognition, Institution Building and Industry Emergence*, Promotors: P.P.M.A.R. Heugens, & Prof. J.P. Cornelissen, Eps-2019-483-S&E, https://repub.eur.nl/pub/118760

Straeter, L.M., *Interpersonal Consumer Decision Making*, Promotors: Prof. S.M.J. van Osselaer & Dr I.E. de Hooge, EPS-2017-423-MKT, https://repub.eur.nl/pub/100819

Stuppy, A., *Essays on Product Quality*, Promotors: Prof. S.M.J. van Osselaer & Dr N.L. Mead. EPS-2018-461-MKT, https://repub.eur.nl/pub/111375

Subaşi, B., *Demographic Dissimilarity, Information Access and Individual Performance*, Promotors: Prof. D.L. van Knippenberg & Dr W.P. van Ginkel, EPS-2017-422-ORG, https://repub.eur.nl/pub/103495

Suurmond, R., *In Pursuit of Supplier Knowledge: Leveraging capabilities and dividing responsibilities in product and service contexts*, Promotors: Prof. J.Y.F Wynstra & Prof. J. Dul. EPS-2018-475-LIS, https://repub.eur.nl/pub/115138

Toxopeus, H.S. *Financing sustainable innovation: From a principal-agent to a collective action perspective*, Promotors: Prof. H.R. Commandeur & Dr. K.E.H. Maas. EPS-2019-458-S&E, https://repub.eur.nl/pub/114018

Tuijn, M. C., *Target the untargeted: essays in unconventional disclosures and policies*, Promotors: Prof. E. Peek & Prof. E.M. Roelofsen, EPS-2020-499-F&A, https://repub.eur.nl/pub/125625

Turturea, R., *Overcoming Resource Constraints: The Role of Creative Resourcing and Equity Crowdfunding in Financing Entrepreneurial Ventures*, Promotors: Prof. P.P.M.A.R Heugens, Prof. J.J.P. Jansen & Dr. I. Verheuil, EPS-2019-472-S&E, https://repub.eur.nl/pub/112859

Valboni, R., *Building Organizational (Dis-)Abilities: The impact of learning on the performance of mergers and acquisitions*, Promotors: Prof. T.H. Reus & Dr A.H.L. Slangen, EPS-2020-407-S&E, https://repub.eur.nl/pub/125226

Vandic, D., *Intelligent Information Systems for Web Product Search*, Promotors: Prof. U. Kaymak & Dr Frasincar, EPS-2017-405-LIS, https://repub.eur.nl/pub/95490

Verbeek, R.W.M., *Essays on Empirical Asset Pricing*, Promotors: Prof. M.A. van Dijk & Dr M. Szymanowska, EPS-2017-441-F&A, https://repub.eur.nl/pub/102977

Visser, T.R. *Vehicle Routing and Time Slot Management in Online Retailing*, Promotors: Prof. A.P.M. Wagelmans & Dr R. Spliet, EPS-2019-482-LIS, https://repub.eur.nl/pub/120772

Vlaming, R. de., *Linear Mixed Models in Statistical Genetics*, Prof. A.R. Thurik, Prof. P.J.F. Groenen & Prof. Ph.D. Koellinger, EPS-2017-416-S&E, https://repub.eur.nl/pub/100428

Vongswasdi, P., *Accelerating Leadership Development: An evidence-based perspective*, Promotors: Prof. D. van Dierendonck & Dr H.L. Leroy, EPS-2020-512-ORG, https://repub.eur.nl/pub/134079

Vries, H. de, *Evidence-Based Optimization in Humanitarian Logistics*, Promotors: Prof. A.P.M. Wagelmans & Prof. J.J. van de Klundert, EPS-2017-435-LIS, https://repub.eur.nl/pub/102771

Wang, R., *Corporate Environmentalism in China*, Promotors: Prof. P.P.M.A.R Heugens & Dr F. Wijen, EPS-2017-417-S&E, https://repub.eur.nl/pub/99987

Wang, R., *Those Who Move Stock Prices*, Promotors: Prof. P. Verwijmeren & Prof. S. van Bekkum, EPS-2019-491-F&A, https://repub.eur.nl/pub/129057

Wasesa, M., *Agent-based inter-organizational systems in advanced logistics operations*, Promotors: Prof. H.W.G.M van Heck, Prof. R.A. Zuidwijk & Dr A. W. Stam, EPS-2017-LIS-424, https://repub.eur.nl/pub/100527

Wessels, C., *Flexible Working Practices: How Employees Can Reap the Benefits for Engagement and Performance*, Promotors: Prof. H.W.G.M. van Heck, Prof. P.J. van Baalen & Prof. M.C. Schippers, EPS-2017-418-LIS, https://repub.eur.nl/pub/99312

Wiegmann, P.M., *Setting the Stage for Innovation: Balancing Diverse Interests through Standardisation*, Promotors: Prof. H.J. de Vries & Prof. K. Blind, EPS-2019-473-LIS, https://repub.eur.nl/pub/114519

Wijaya, H.R., *Praise the Lord!: Infusing Values and Emotions into Neo-Institutional Theory*, Promotors: Prof. P.P.M.A.R. Heugens & Prof. J.P. Cornelissen, EPS-2019-450-S&E, https://repub.eur.nl/pub/115973

Williams, A.N., *Make Our Planet Great Again: A Systems Perspective of Corporate Sustainability*, Promotors: Prof. G.M. Whiteman & Dr. S. Kennedy, EPS-2018-456-ORG, https://repub.eur.nl/pub/111032

Witte, C.T., *Bloody Business: Multinational investment in an increasingly conflict-afflicted world*, Promotors: Prof. H.P.G. Pennings, Prof. H.R. Commandeur & Dr M.J. Burger, EPS-2018-443-S&E, https://repub.eur.nl/pub/104027

Wu, J., *A Configural Approach to Understanding Voice Behavior in Teams*, Promotors: Prof. D.L. van Knippenberg & Prof. S.R. Giessner, EPS-2020-510-ORG, https://repub.eur.nl/pub/132184

Ye, Q.C., *Multi-objective Optimization Methods for Allocation and Prediction*, Promotors: Prof. R. Dekker & Dr Y. Zhang, EPS-2019-460-LIS, https://repub.eur.nl/pub/116462

Yuan, Y., *The Emergence of Team Creativity: a social network perspective*, Promotors: Prof. D. L. van Knippenberg & Dr D. A. Stam, EPS-2017-434-ORG, https://repub.eur.nl/pub/100847

Zhang, Q., *Financing and Regulatory Frictions in Mergers and Acquisitions*, Promotors: Prof. P.G.J. Roosenboom & Prof. A. de Jong, EPS-2018-428-F&A, https://repub.eur.nl/pub/103871

Accurate judgment, adaptive decision making, and ability to find insightful solutions to challenging problems are some of the key qualities organizations aspire to instill in their members. In the service of this goal, designers of management systems draw on a variety of approaches, ranging from formal contracts to subtle "nudges." At the core of many of these approaches is the idea of *accountability*, the use of external scrutiny to bring about desired behavior and outcomes. How should accountability be implemented to elicit optimal performance from individuals? Does scrutinizing the *process* the accountable individual is using tend to yield more accurate judgments and decisions than focusing strictly on the *outcome*? Does the answer depend on the nature of the decision environment? Does the answer hold in the domain of problems that call for insightful solutions? What are the cognitive and affective mechanisms that mediate the differences in performance between *process accountability* and *outcome accountability*? Are these effects robust or do they depend on currently insufficiently understood moderating factors? This dissertation reports the results of a series of behavioral experiments designed to shed light on these and related questions.

ERiM

The Erasmus Research Institute of Management (ERIM) is the Research School (Onderzoekschool) in the field of management of the Erasmus University Rotterdam. The founding participants of ERIM are the Rotterdam School of Management (RSM), and the Erasmus School of Economics (ESE). ERIM was founded in 1999 and is officially accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW). The research undertaken by ERIM is focused on the management of the firm in its environment, its intra- and interfirm relations, and its business processes in their interdependent connections.

The objective of ERIM is to carry out first rate research in management, and to offer an advanced doctoral programme in Research in Management. Within ERIM, over three hundred senior researchers and PhD candidates are active in the different research programmes. From a variety of academic backgrounds and expertises, the ERIM community is united in striving for excellence and working at the forefront of creating new business knowledge.

ERiM

ERIM PhD Series
Research in Management

Erasmus University Rotterdam (EUR)
Erasmus Research Institute of Management
Mandeville (T) Building
Burgemeester Oudlaan 50
3062 PA Rotterdam, The Netherlands

P.O. Box 1738
3000 DR Rotterdam, The Netherlands
T +31 10 408 1182
E info@erim.eur.nl
W www.erim.eur.nl