CERGE-EI
Center for Economic Research and Graduate Education – Economics Institute

Essays in Information Economics

Rastislav Rehák

Dissertation

Prague 2023

Dissertation Committee

Filip Matějka (CERGE-EI; chair)

Jakub Steiner (CERGE-EI and University of Zurich)

Jan Zápal (CERGE-EI)

Referees

David Walker-Jones (University of Surrey)

Dong Wei (University of California, Santa Cruz)

Contents

Acknowledgments

I would like to take this opportunity to express my sincere gratitude to those who have supported me throughout my academic journey. First and foremost, I am grateful to my supervisor Filip Matějka, for his invaluable guidance, encouragement, and support throughout the research process. He is someone I greatly admire and consider a role model, both for his intellect and his character. I would also like to express my gratitude and admiration to Jan Zápal, Ole Jann, Jakub Steiner, Stanislav Anatolyev, Michal Bauer, and Julie Chytilová. They have been an inspiration for me and their insightful lectures, comments, and feedback have been instrumental in shaping my work.

I am also indebted to my co-authors, colleagues, and friends, in particular Maxim Senkov, Darya Korlyakova, Sona Badalyan, Michal Hakala, Martin Štrobl, Pavel Ilinov, Pavel Kocourek, Artem Razumovskii, Vladimír Novák, and Andrei Matveenko. They made my PhD studies very enjoyable and I have always felt comfortable discussing my thoughts with them, no matter how crude, erroneous, and weird.

I am grateful to the whole CERGE-EI community—faculty members, students, and staff—for providing amazing support, a cozy environment, and a great program. My time spent at CERGE-EI has been probably the most amazing time of my life so far.

Financial support by Charles University (GAUK project No. 666420), the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreements No. 101002898 and No. 770652), the H2020-MSCA-RISE project GEMCLIME-2020 GA No. 681228, the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 870245, and the Lumina Quaeruntur fellowship (LQ300852101) of the Czech Academy of Sciences is gratefully acknowledged. Their generosity has enabled me to pursue my academic goals and aspirations.

I would like to extend my thanks to Pietro Ortoleva, Tomasz Strzalecki, and Xavier Gabaix for their hospitality and support during my research stays at Harvard and Princeton. I also want to thank João Thereze, Francesco Fabbri, Lukas Freund, Minghao Zou,

Carol Shou, Paola Moscariello, and Andrew Ferdowsian for being such a fun, welcoming, and inspiring crowd during my stay at Princeton.

Last but not least, I would like to express my heartfelt gratitude to my family for their unwavering support and encouragement. Their love, patience, and understanding have been a constant source of strength. My wife, Barbora Rehák Bučková, has been a special source of strength in this academic extravaganza.

In conclusion, I am grateful to all those who have supported me in one way or another throughout my academic journey. Your contributions have been invaluable.

Prague, Czech Republic — Rastislav Rehák
July 2023

Abstract

In the first chapter, we study informational dissociations between decisions and decision confidence. We explore the consequences of a dual-system model: the decision system and confidence system have distinct goals, but share access to a source of noisy and costly information about a decision-relevant variable. The decision system aims to maximize utility while the confidence system monitors the decision system and aims to provide good feedback about the correctness of the decision. In line with existing experimental evidence showing the importance of post-decisional information in confidence formation, we allow the confidence system to accumulate information after the decision. We aim to base the post-decisional stage (used in descriptive models of confidence) in the optimal learning theory. However, we find that it is not always optimal to engage in the second stage, even for a given individual in a given decision environment. In particular, there is scope for post-decisional information acquisition only for relatively fast decisions. Hence, a strict distinction between one-stage and two-stage theories of decision confidence may be misleading because both may manifest themselves under one underlying mechanism in a non-trivial manner.

In the second chapter, we study a Bayesian persuasion model in which the state space is finite, the sender and the receiver have state-dependent quadratic loss functions, and their disagreement regarding the preferred action is of arbitrary form. This framework enables us to focus on the understudied sender's trade-off between the informativeness of the signal and the concealment of the state-dependent disagreement about the preferred action. In particular, we study which states are pooled together in the supports of posteriors of the optimal signal. We provide an illustrative graph procedure that takes the form of preference misalignment and outputs potential representations of the state-pooling structure. Our model provides insights into situations in which the sender and the receiver care about two different but connected issues, for example, the interaction of a political advisor who cares about the state of the economy with a politician who cares about the political situation.

In the third chapter, we focus on communication among hiring team members and doc-

ument the existence of discrimination in the disclosure of information about candidates. In particular, we conduct an online experiment with a nationally representative sample of Czech individuals who act as human resource assistants and hiring managers in our online labor market. The main novel feature of our experiment is the monitoring of information flow between human resource assistants and hiring managers. We exogenously manipulate candidates' names to explore the causal effects of their gender on information that assistants select for managers. Our findings reveal that assistants disclose more information about family and less information about work for female candidates than for male candidates. An in-depth analysis of types of information disclosed suggests that gender stereotypes play an important role in this disclosure discrimination.

Introduction

The overarching theme that ties together all three chapters of this dissertation is the role of information. Information is a crucial component of learning, decision-making, and effective communication, and this dissertation sheds light on the diverse ways in which it can be utilized. In the first chapter, we delve into a dynamic model of costly information acquisition, which theorizes about decision confidence formation. The second chapter examines a model of strategic information design. Finally, the third chapter presents an experimental investigation into whether HR assistants exhibit discriminatory practices in the information they disclose about job candidates based on the candidates' gender.

In the first chapter, we ask when acquiring additional information after a decision is optimal for refining decision confidence. By doing so, we provide a normative foundation for post-decisional information acquisition featured in descriptive models in cognitive science. These models assume post-decisional information acquisition after each decision to fit experimental data; those data suggest that people actively acquire information about the displayed options even after making decisions. However, we show that the mechanism of post-decisional information acquisition may be more nuanced. We show that in every decision environment, it is not optimal to engage in post-decisional information acquisition for sufficiently slow decisions, while decisions made quickly may lead to it. These findings unify and clarify the earlier cognitive science literature by extending the

statistical sequential sampling literature.

In the second chapter, we tackle the problem of characterization of optimal information design. Specifically, we study a particular instance of the general Bayesian persuasion model that has received little attention in the literature—one that allows for an arbitrary disagreement about the preferred action of the sender and the receiver in each state of the world. In our setup, the state space is finite, and the sender and the receiver have quadratic loss functions from the action the receiver implements relative to a state-dependent preferred action. This framework enables us to focus on the understudied sender's trade-off between the informativeness of the information structure and the concealment of the state-dependent disagreement about the preferred action. Specifically, we analyze which states are strategically pooled together (in the supports of posteriors of the optimal information structure). The state-pooling structure may shed light on "language" used in strategic situations in which the sender and the receiver care about two different but connected issues, e.g., a political advisor who cares about the state of the economy and a politician who cares about the political situation. We provide an illustrative graph procedure that takes the form of preference misalignment and outputs potential representations of the state-pooling structure.

In the third chapter, we investigate communication among members of a hiring team and uncover gender-based discrimination in the disclosure of information about job candidates. We employ a nationally representative sample of Czech individuals who act as human resource assistants and hiring managers in our online labor market. The unique aspect of our experiment is that we monitor the information flow between assistants and managers, and manipulate gender of job candidates' names to explore the causal effects of gender on the type of information that assistants share with managers. Our findings reveal that assistants disclose more information about family and less about work responsibilities for female candidates than for male candidates. A closer examination of the types of information shared and further heterogeneity analyses suggest that gender stereotypes are a significant factor in this discriminatory behavior.

Chapter 1

Sequential Sampling Beyond Decisions? A Normative Model of Decision Confidence

Abstract

Rastislav Rehák[1]

We study informational dissociations between decisions and decision confidence. We explore the consequences of a dual-system model: the decision system and confidence system have distinct goals, but share access to a source of noisy and costly information about a decision-relevant variable. The decision system aims to maximize utility while

[1]This chapter is based on Rehák, R. (2022) "Sequential Sampling Beyond Decisions? A Normative Model of Decision Confidence," CERGE-EI Working Paper Series No. 739. This project was supported by Charles University GAUK project No. 666420 and by the H2020-MSCA-RISE project GEMCLIME-2020 GA No. 681228. This paper is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 870245. This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreements No. 101002898 and No. 770652). Parts of this paper were developed and written during my research stays at Harvard University and Princeton University. I am very grateful to Filip Matějka for guidance and support, to Alex Bloedel, Tom Griffiths, Pavel Kocourek, Xiaosheng Mu, Pietro Ortoleva, Maxim Senkov, Milan Ščasný, João Thereze, Can Urgun, Ansgar Walther, and Weijie Zhong for very useful discussions, and to numerous other discussants for insights and comments.

the confidence system monitors the decision system and aims to provide good feedback about the correctness of the decision. In line with existing experimental evidence showing the importance of post-decisional information in confidence formation, we allow the confidence system to accumulate information after the decision. We aim to base the post-decisional stage (used in descriptive models of confidence) in the optimal learning theory. However, we find that it is not always optimal to engage in the second stage, even for a given individual in a given decision environment. In particular, there is scope for post-decisional information acquisition only for relatively fast decisions. Hence, a strict distinction between one-stage and two-stage theories of decision confidence may be misleading because both may manifest themselves under one underlying mechanism in a non-trivial manner.

Keywords: Decision, Confidence, Sequential Sampling, Optimal Stopping
JEL Codes: C11, C41, C44, D11, D83, D91

1.1 Introduction

Decision confidence is a subjective assessment of one's own decision quality—a belief that a decision is correct. It is a manifestation of one's metacognitive abilities, which has been an area of great interest in cognitive science. However, it has received much less attention in economics even though it is relevant in many situations of economic interest.[2] Moreover, decision confidence provides cheap additional information about people's preferences and judgements, which relates to the proposal of Krajbich, Oud, and Fehr (2014) to use previously neglected measures.

We jointly model decisions, decision times, confidence judgements, and interjudgement times.[3] Such data are commonly gathered in the two-alternative choice-followed-by-

[2] We list some areas and literature that provide economic motivation for interest in decision confidence in Appendix 1.F.

[3] *Confidence judgements* can be recorded on a slider between 0 and 1, with 1 corresponding to being certain about the decision and 0 to recognizing an error with certainty. *Decision times* (also called 'reaction times' or RT) measure the time between the start of the visual display of stimulus and the recording of the decision. *Interjudgement times* (also denoted RT2) measure the time between the recording of the decision and the recording of the confidence judgement.

confidence experimental paradigm.[4] For example, a participant may be asked to choose which cloud on the screen (left or right) contains more dots and, after making the decision, to indicate on a slider between 0 and 1 her confidence in the decision.

In cognitive science, several descriptive[5] models have been developed to capture this kind of data. We relate directly to two prominent models in this literature: the Two-Stage Dynamic Signal Detection (2DSD) model of Pleskac and Busemeyer (2010) and the Collapsing Confidence Boundary (CCB) model of Moran, Teodorescu, and Usher (2015). We adopt the idea of (potential) post-decisional evidence accumulation for confidence judgement formation (Yeung and Summerfield 2012; Fleming and Daw 2017). The primary role of these models is to provide a flexible framework that is able to capture a multitude of empirical patterns and thus enable us to summarize and think about those patterns in a succinct way. However, these models are heuristic in the sense that they are not derived from first principles and the objects characterizing the behavior are not tied to the primitives of the decision environment. To be able to conduct counterfactual analyses and theorize about the drivers of decision confidence, normative models are needed.

At the normative spectrum of the literature, we follow closely the model of Fudenberg, Strack, and Strzalecki (2018). However, this and similar normative models (Wald 1947; Tajima, Drugowitsch, and Pouget 2016) were developed to model primarily decisions, but not decision confidence. On one hand, they feature belief confidence,[6] so one can define decision confidence in these models as the belief confidence in the chosen option upon

[4]The largest available source of data is the Confidence Database (Rahnev et al. 2020).

[5]We use the nomenclature of *normative* and *descriptive* models in the sense of Baron (2012). In particular, descriptive models "try to explain how people make judgments and decisions" (Baron 2012, p. 1); they are data-driven. On the other hand, normative models serve as "standards for evaluation" and "[t]hey must be justified independently of observations of people's judgments and decisions, once we have observed enough to define what we are talking about." (Baron 2012, p. 1). Hence, normative models capture our understanding of a situation, formulate a problem associated with that situation, and assume an optimal solution to the problem to derive behavior.

[6]To clarify the terminology, *belief confidence* (in an option) is accessible at any point of deliberation. However, in this paper, we are modeling *decision confidence*, which we define as a subjective assessment of one's own decision quality. In particular, this definition requires implicitly that (i) a decision is made (i.e., there is no decision confidence without a decision) and (ii) decision confidence is a committed judgement (i.e., the endpoint of deliberation, not a running variable). This distinction mirrors the distinction between "confidence" and "certainty" advocated by Pouget, Drugowitsch, and Kepecs (2016).

stopping. On the other hand, such decision confidence is incompatible with the observed ability of people to recognize their own errors even without explicit feedback (Yeung and Summerfield 2012; Fleming and Daw 2017) and the importance of post-decisional evidence in confidence formation (Moran, Teodorescu, and Usher 2015). Hence, an extension of these models is needed to account for the whole mechanism behind decision confidence.

In this paper, we develop a normative dynamic model of decision confidence. At Marr's (1982) computational level, we posit that a decision maker employs two systems with distinct goals—a *decision system* and a *confidence system*. The decision system aims to choose the best option, i.e., it maximizes expected utility. The confidence system aims to monitor the decision system and provide good feedback on its performance. We assume that confidence acts as a substitute for explicit feedback (thus, it plays a key role in situations when explicit feedback is not (immediately) available); the confidence system minimizes the mean-squared error (MSE) of confidence relative to the perfect feedback indicator of (in)correctness of the decision. An implication of this assumption is that decision confidence is the posterior probability of being correct, in accordance with the Bayesian confidence hypothesis (Pouget, Drugowitsch, and Kepecs 2016). Finally, the two separate systems have access to a common source of costly and noisy evidence about the values of the options and resolve the speed-performance trade-off optimally. Naturally, we assume that the confidence system can continue evidence accumulation beyond decisions.

We are interested especially in dissociations between decision performance and metacognition. Specifically, we ask when decision and decision confidence should be based on the same evidence. Stated differently, we allow for two-stage confidence formation and we ask when it is optimal to use only one stage. This amounts to the comparison of the optimal stopping regions for evidence accumulation of the decision stopping problem and the unconstrained confidence stopping problem.[7]

Our main analytical result is a closed-form bound on the unconstrained confidence stopping time of the evidence process.[8] Together with the unboundedness of the decision stopping time, this implies that (relatively)[9] slow decisions will lead to so called *one-*

[7]We express both stopping algorithms in the space of the evidence process (as opposed to the space of posterior expected values, for example).

[8]The original confidence stopping problem is constrained by the decision stopping time.

[9]In short, "relatively" means relative to a decision environment. Intuitively, choosing a house in ten minutes is fast, but choosing an apple for a snack in ten minutes is slow. We discuss in detail the meaning

stage confidence, which is a situation when the confidence is based on the same evidence as the decision. Therefore, there is scope for *two-stage confidence*, which is a situation when the confidence is based on more evidence than the decision, only for fast decisions. Consequently, there is space for error monitoring only for the fast decisions. However, we demonstrate numerically that under some parameters, not all fast decisions must lead to two-stage confidence. Surprisingly, it may happen that the fastest decisions (together with the slow decisions) lead to one-stage confidence, while only intermediately fast decisions lead to two-stage confidence. Finally, an intuitive result is that it is only under low cost of time and/or strong preference for good confidence that there is room for two-stage confidence at all.

We contribute to two main strands of literature. First, we build on the analysis of Chernoff's (1961) problem of sequential testing of the sign of the normal drift of a Brownian motion (Zhitlukhin and Muravlev 2013; Fudenberg, Strack, and Strzalecki 2018). However, our confidence stopping problem is, to the best of our knowledge, a novel problem. Second, we contribute to the first long-term goal for the field of metacognition formulated by Rahnev et al. (2021, p. 6)—development of detailed models of visual metacognition. To the best of our knowledge, we are the first to propose a *dynamic normative model* of decision confidence.

Our main contribution to the discussion about one-stage vs. two-stage theories of decision confidence is that a strict distinction between them (Moran, Teodorescu, and Usher 2015) might be misleading because both may manifest themselves under one underlying mechanism, even within one individual in a given controlled decision environment. Moreover, our approach allows us to predict how an individual might change the modes of decision confidence formation under different circumstances. Hence, our model speaks to both intra- and inter-individual differences in the formation of decision confidence.

In a complementary work, Fleming and Daw (2017) propose a Bayesian framework for grounding a discussion about a related aspect of metacognitive computation—whether decision and confidence are informed by the same signal or different but correlated signals. In their framework, our model falls into the category of "postdecisional" models (Pleskac and Busemeyer 2010; Moran, Teodorescu, and Usher 2015), in which a single process informs both decision and decision confidence. We are thus leaving aside their proposed

---

of "relatively" (fast/slow decisions) in Section 1.3.

"second-order" architecture that allows decision and decision confidence to be informed by distinct but correlated processes. However, our model has a feature reminiscent of the "second-order" architecture: the goals of the decision and confidence systems are distinct. Neverthless, Fleming and Daw (2017) do not postulate an explicit goal for the confidence system (for the decision system, it is implicitly expected utility maximization). Moreover, they work in a static environment and are concerned with a high-level structure of confidence computation, while we focus on *procedural* details of the *optimal* evidence accumulation.[10]

The question we ask—when is it optimal to gather additional evidence after the decision?—is similar to the question studied by the literature about metacognitive control (Schulz, Fleming, and Dayan 2021; Boldt, Blundell, and De Martino 2019; Desender, Boldt, and Yeung 2018). In a typical experimental paradigm in this literature, participants are asked to decide whether to obtain additional information after the first decision/stimulus presentation, but they are given an explicit motivation for doing so, e.g., a revision of the initial decision or a subsequent related decision. Moreover, the stimulus presentation is often not under full control of the participant. In contrast, we aim to contribute primarily to the literature about metacognitive monitoring, i.e., we are interested in how decision confidence arises rather than how it is used to control subsequent behavior. In particular, sampling beyond decisions in our setup leads to formation of decision confidence, while in the control literature, the roles are reversed—decision confidence is used to decide about additional sampling for a specified goal.

1.2 Model

Our model consists of two separate systems—a decision system and a confidence system. The systems have distinct goals, but they have access to a common evidence process and sampling is costly. The decision system's goal is to maximize expected utility. The confidence system's goal is to accurately assess the choice (metacognitive monitoring).

[10] Fleming and Daw (2017) recognize the importance of modeling the dynamics in their footnote 1, p. 94.

1.2.1 Information technology

The agent chooses between options l and r, which can bring her utilities $\theta^l \in \mathbb{R}$ and $\theta^r \in \mathbb{R}$, respectively. The agent does not know the true utilities a priori, but she can sample evidence about them to make an informed decision.[11]

The agent's object of interest is a sufficient statistic about the options' utilities θ^l and θ^r, which we denote by θ. We leave the functional form of $\theta(\theta^l, \theta^r)$ open, but it is supposed to measure a dissociation between the two options, e.g., their difference. We assume that the agent cares about the sign of θ and that she has a normal prior about it

$$N(X_0, 2\sigma_0^2).^{12} \tag{1.1}$$

The agent can learn about the true θ by observing a continuous signal $\{Z_t\}_{t\geq 0}$

$$Z_t = \theta t + \alpha\sqrt{2}B_t, t \geq 0, \tag{1.2}$$

where $\{B_t\}_{t\geq 0}$ is a standard Brownian motion independent of θ and parameter $\alpha > 0$ captures the strength of the noise. The agent pays constant flow cost $c > 0$ for observing this process.

We denote the information up to time t by $\mathcal{F}_t$ (formally, on our probability space $(\Omega, \mathcal{F}, \mathsf{P})$, we have a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ generated by process $\{Z_t\}_{t\geq 0}$). We denote the posterior mean and variance about θ at time t by $X_t = \mathrm{E}[\theta|\mathcal{F}_t]$ and $\sigma_t^2 = \mathrm{var}(\theta|\mathcal{F}_t)$, respectively.

---

[11] We interpret the evidence process broadly as a subjective process of introspection, recollection of past experiences and information, assessment of visual and other stimuli about the options, etc.

[12] This allows for interpretation of θ as the difference $\theta^l - \theta^r$ of two jointly normal utilities (θ^l, θ^r), as in Fudenberg, Strack, and Strzalecki (2018) (we try to adhere to their notation for a more transparent connection to their paper—that is also the reason to write variance in the form $2\sigma_0^2$). Arguably, this is all the agent should care about when deciding which option is better. This interpretation is more suited to value-based decisions for which θ^l and θ^r already refer to variables internal to the decision maker. However, we can also interpret θ as $\ln(\theta^l/\theta^r)$—the natural logarithm of the ratio of two independent log-normal magnitudes θ^l and θ^r. This second interpretation is more suited to perceptual decisions for which θ^l and θ^r refer to objective stimuli manipulated by the experimenter. In particular, the second interpretation can capture the Weber-Fechner law (e.g., distinguishing between boxes with 95 and 100 dots is more difficult than distinguishing between boxes with 5 and 10 dots).

1.2.2 Decision system

The decision system's goal is to design a decision rule and a sampling rule to support that decision.

Conditional on stopping at time t, it is optimal to choose an option according to posterior expectation X_t. Hence, the optimal decision rule is $\text{sgn}(X_t)$ with the understanding that a positive sign indicates the choice of option l and a negative sign indicates the choice of option r.

Taking this decision rule into account, the decision system designs a stopping time in order to maximize the expected probability of being correct, weighted by the importance of the decision minus the cost of sampling time. We capture the last sentence formally. First, an $\mathcal{F}_t$-stopping time is a rule that, for each path of the evidence process $(Z_t(\omega))_{t\geq 0}$, $\omega \in \Omega$, prescribes when to stop its observation in a manner consistent with the arrival of information about which path is actually realized (modeled by filtration $\{\mathcal{F}_t\}_{t\geq 0}$). Formally, $\tau \colon \Omega \to [0,\infty]$ is a random variable such that $\{\tau \leq t\} \in \mathcal{F}_t \ \forall t \geq 0$ and $\mathsf{P}(\tau < \infty) = 1$. We denote by $\mathcal{T}$ the set of all $\mathcal{F}_t$-stopping times. Second, the probability of being correct is the probability that the estimated sign of θ is equal to its actual sign, $\text{sgn}(X_t) = \text{sgn}(\theta)$. Finally, the importance of the decision is captured by the absolute value of θ, i.e., choosing correctly from two close options is less important than choosing correctly when one option is substantially inferior to the other. Hence, the decision system faces the stopping problem

$$\sup_{\tau\in\mathcal{T}} \mathrm{E}\left[|\theta| \mathbb{1}\{\text{sgn}(X_\tau) = \text{sgn}(\theta)\} - c\tau\right].^{13,14} \tag{1.3}$$

1.2.3 Confidence system

The agent can continue to accumulate evidence $\{Z_t\}_{t\geq 0}$ even after the end of the decision stage τ in order to refine her degree of confidence in her decision. By accumulating

[13] In cases of multiple optimal stopping times, we select the minimal optimal stopping time. This is assumed also for the confidence stopping problem.

[14] In Appendix 1.A.1, we comment on this objective and its connection to other formulations of the problem that appeared in Chernoff (1961), Zhitlukhin and Muravlev (2013), and Fudenberg, Strack, and Strzalecki (2018).

evidence in the confidence stage, the agent pays a constant flow cost $\bar{c} > 0$. The decision system's goal is to design a confidence judgement and a sampling rule to support that judgement.

In accordance with the Bayesian confidence hypothesis (Pouget, Drugowitsch, and Kepecs 2016), we define *decision confidence* as the probability of having made the correct decision.[15] Hence, decision confidence at time $t \geq \tau$ is

$$conf_t = \mathsf{P}\left(\operatorname{sgn}(X_\tau) = \operatorname{sgn}(\theta) \middle| \mathcal{F}_t\right). \tag{1.4}$$

The agent decides when to stop the confidence stage in order to minimize the MSE loss from incorrectly assessing the (objective) (in)correctness of the decision and the additional cost of evidence accumulation in the confidence stage

$$\inf_{\tau_c \in \mathcal{T} \text{ s.t. } \tau_c \geq \tau} \mathrm{E}\left[\left(conf_{\tau_c} - \mathbb{1}\{\operatorname{sgn}(X_\tau) = \operatorname{sgn}(\theta)\}\right)^2 + \bar{c}(\tau_c - \tau)\right]. \tag{1.5}$$

Note that flow cost $\bar{c}$ is likely to be different from c because it is expressed in different units: while c captures forgone utils during decision making, $\bar{c}$ captures forgone utils during confidence formation relative to lost utils due to misplaced (inaccurate) confidence. Formally, the most natural interpretation of $\bar{c}$ is that it is cost of time c relative to the importance of accurate confidence denoted by γ, $\bar{c} = \frac{c}{\gamma}$. However, c and $\bar{c}$ can even be manipulated independently in experiments, e.g., a "time-pressure-on-choice" manipulation can be understood as an increase in c without a change in $\bar{c}$.

1.3 (Non)occurrence of post-decisional information acquisition

Our main insight is that post-decisional information acquisition after each decision, which is assumed in some prominent models in cognitive science (Pleskac and Busemeyer 2010; Moran, Teodorescu, and Usher 2015), is not supported in our model. In every decision environment, it is not optimal to engage in post-decisional information acquisition for

[15] This definition of decision confidence can also be justified as the optimal assessment under the mean-squared error (MSE) loss from incorrectly assessing the (objective) (in)correctness of the decision.

sufficiently slow decisions. On the other hand, there is some scope for post-decisional information acquisition for relatively fast decisions. We state this result formally in the following theorem.

Theorem 1.1. *Let* $T_c = \max\{\frac{1}{2\pi\bar{c}} - \frac{\alpha^2}{\sigma_0^2}, 0\}$ *and denote by* τ^* *and* τ_c^* *the minimal optimal decision and confidence stopping times, respectively. Then* $\mathsf{P}(\tau^* > T_c, \tau_c^* = \tau^*) = \mathsf{P}(\tau^* > T_c) > 0$.

Proof. The theorem follows from Lemma 1.A.3 and Lemma 1.A.4 in the Appendix. □

This theorem provides an explicit bound T_c delineating the "sufficiently slow" decisions. Then it states that there is always (i.e., for all decision environments) a positive mass of those decisions—$\mathsf{P}(\tau^* > T_c) > 0$—and that they almost surely do not lead to post-decisional information acquisition—$\mathsf{P}(\tau^* > T_c, \tau_c^* = \tau^*) = \mathsf{P}(\tau^* > T_c)$.

The bound T_c in Theorem 1.1 varies intuitively with the parameters of the decision environment: it increases with lower cost of time relative to the importance of precise confidence $\bar{c}$, lower noisiness of the evidence process α, and higher prior variance σ_0^2. Prior variance is higher for the choice of a house than for the choice of a snack, so choosing a house in ten minutes might be considered fast, while choosing a snack in ten minutes might be considered slow, for example. Hence, "fast" decisions determined by this threshold naturally bear different meaning in different contexts. Therefore, we emphasize that the definition of "fast" and "slow" decisions is relative to the decision environment—it is supposed to capture a qualitative distinction between decisions in terms of timing for a given decision environment, but it does not imply that there is a universal fixed time threshold that could classify decisions as fast or slow.

Notice that the cost of time from the decision problem c does not appear in the formula for T_c. This is because the objective of the confidence system (1.5) does not depend explicitly on c. Moreover, the confidence system only evaluates the outcomes of the decision system (choice and decision time), but does not strategically respond to them by adapting its inherent criteria for satisfactory evidence.

Theorem 1.1 stems from the dissociation between the goals of the decision and confidence systems. While both systems follow the same belief process, the confidence system does not care about a particular position of this process, only about its precision. On the other hand, the decision system cares about one particular position of the belief process—the

point of indifference—because of the discreteness of the action space and non-smooth change in expected utility at this point.

We can gain intuition about the functioning of these two systems by drawing an analogy to measurement of weight on a mechanical scale with a needle. The confidence system functions as an impartial reader of weight that cares about stabilizing random movements of the needle around some value just enough to obtain a reliable (albeit still imprecise) practical sense of her weight. On the other hand, the decision system functions as a wrestler who aims at attaining maximum weight while staying in his weight category. For this wrestler, the needle too close to the threshold calls for more measurement time in order to resolve whether there is a need for weight loss. Hence, signals on different sides of the threshold have very different consequences for the wrestler. However, all the wrestler cares about are the consequences for his dietary regime and the precision plays only an ancillary role. Therefore, at the beginning of his season, when his weight is not fine-tuned yet, he might be satisfied with a coarse measurement giving him only the necessary direction for action, which is in contrast with the impartial reader.

1.4 Analysis

In this section, we provide an outline of the analysis leading to Theorem 1.1 and graphical illustration based on numerical solutions, to give more insight into the mechanics implied by the model.

Our research question leads to the comparison of the solutions of the optimal stopping problems of the decision and confidence system. Since the stopping problem of the decision system has been studied elsewhere (in particular, by Fudenberg, Strack, and Strzalecki [2018]), we focus on the confidence stopping problem.

1.4.1 Reformulation of the confidence problem

By Lemma 1.A.1 in the Appendix, the beliefs about θ at time t are normal

$$N(X_t, \sigma_t^2) \tag{1.6}$$

with mean

$$X_t = \frac{\sigma_0^{-2} X_0 + \alpha^{-2} Z_t}{\sigma_0^{-2} + \alpha^{-2} t} \tag{1.7}$$

and variance

$$\sigma_t^2 = \frac{2}{\sigma_0^{-2} + \alpha^{-2} t}. \tag{1.8}$$

Hence, we can express the confidence explicitly

$$conf_t = \Phi\left(\frac{X_t}{\sigma_t}\right) \mathbb{1}\{X_\tau \geq 0\} + \Phi\left(-\frac{X_t}{\sigma_t}\right) \mathbb{1}\{X_\tau < 0\}, \tag{1.9}$$

where Φ is the CDF of the standard normal distribution.

Since

$$\begin{aligned} &\mathrm{E}\left[\left(conf_{\tau_c} - \mathbb{1}\{\mathrm{sgn}(X_\tau) = \mathrm{sgn}(\theta)\}\right)^2\right] \\ &\quad = \mathrm{E}\left[\mathrm{E}\left[\left(conf_{\tau_c} - \mathbb{1}\{\mathrm{sgn}(X_\tau) = \mathrm{sgn}(\theta)\}\right)^2 \middle| \mathcal{F}_{\tau_c}\right]\right] \\ &\quad = \mathrm{E}\left[\mathrm{var}\left(\mathbb{1}\{\mathrm{sgn}(X_\tau) = \mathrm{sgn}(\theta)\} \middle| \mathcal{F}_{\tau_c}\right)\right] \\ &\quad = \mathrm{E}\left[conf_{\tau_c}(1 - conf_{\tau_c})\right], \end{aligned} \tag{1.10}$$

we can use (1.9) to rewrite the confidence-stage objective function as

$$\mathrm{E}\left[\Phi\left(\frac{X_{\tau_c}}{\sigma_{\tau_c}}\right) \Phi\left(-\frac{X_{\tau_c}}{\sigma_{\tau_c}}\right) + \bar{c}\tau_c\right] - \bar{c}\mathrm{E}\left[\tau\right]. \tag{1.11}$$

In the reformulation of the confidence objective (1.11), we can focus only on the first part because $\bar{c}\mathrm{E}\left[\tau\right]$ is irrelevant for the choice of τ_c. Moreover, the first part in (1.11) does not feature any τ elements. This inherent independence of the confidence objective on the decision and its timing stems from (i) symmetry and (ii) the strong Markovian property: (i) no matter the decision, the confidence system still cares about both parts of the beliefs, i.e., whether $\theta > 0$ or $\theta < 0$; (ii) the confidence system cares only about the most recent position of the evidence particle. Intuitively, there is no bias toward (dis)confirming the decision—the confidence system acts as an impartial observer who is equally satisfied with a given level of conviction for the conclusion that the decision system is right or wrong. Moreover, it is immaterial for the confidence system to know what evidence the decision system was acting upon (in particular, the amount of evidence

τ); it simply takes that evidence as given and (potentially) builds on it.[16]

Based on these simplifications, we can gain insight into the optimal constrained confidence stopping time in problem (1.5) by studying the auxiliary unconstrained confidence stopping problem

$$\inf_{\tau' \in \mathcal{T}} \mathrm{E}\left[\Phi\left(\frac{\sigma_0^{-2}X_0 + \alpha^{-2}Z_{\tau'}}{\sqrt{2(\sigma_0^{-2}+\alpha^{-2}\tau')}}\right)\Phi\left(-\frac{\sigma_0^{-2}X_0 + \alpha^{-2}Z_{\tau'}}{\sqrt{2(\sigma_0^{-2}+\alpha^{-2}\tau')}}\right) + \bar{c}\tau'\right]. \tag{1.12}$$

Due to the Markovian structure of problem (1.12) (see Appendix 1.A.3), the solution of this problem boils down to finding its continuation region in the (t,z)-space $(= [0,\infty)\times\mathbb{R})$. Specifically, we are looking for a set $C_C \subseteq [0,\infty)\times\mathbb{R}$ such that the complement of C_C is closed and the stopping time $\tau'^* = \inf\{t \geq 0\colon (t, Z_t) \notin C_C\}$ is the (minimal) optimal stopping time in (1.12).

If we denote by C_D the continuation region of the decision stopping problem (1.3) in the (t,z)-space,[17] the continuation region of the original confidence stopping problem (1.5) will be $C_D \cup C_C$. We are interested especially in the analysis of the regions $C_C \backslash C_D$ and $C_D \backslash C_C$, which characterize when it is optimal to continue accumulating evidence beyond the decision stage and when it is optimal to stop immediately after the decision stage, respectively.

1.4.2 Qualitative insights from numerical solutions

In Figure 1.1, we depict the numerically computed decision and confidence stopping boundaries (∂C_D and ∂C_C, respectively) for an actual individual with estimated parameters $c = 0.02, \alpha = 2, \sigma_0 = 1.8, X_0 = 0$[18] and a hypothetical value of $\bar{c} = 0.007$. As is illustrated by this figure, the decision stopping boundary is a pair of barriers collaps-

[16] Our model is applicable to other situations of economic interest, which can provide further intuition for the relation between the decision and the confidence systems described in the previous paragraph. For example, we can think of a CEO of a company who acts as the decision system and a hired external auditor who acts as the confidence system. Similarly, we can think of a politician who acts as the decision system and an unbiased expert/journalist who acts as the confidence system.

[17] Fudenberg, Strack, and Strzalecki (2018) characterize C_D in their Theorem 4 and footnote 22.

[18] See Subject 45 in Table 4 in the Online Appendix of Fudenberg, Strack, and Strzalecki (2018) (the value of $X_0 = 0$ is imposed in their work).

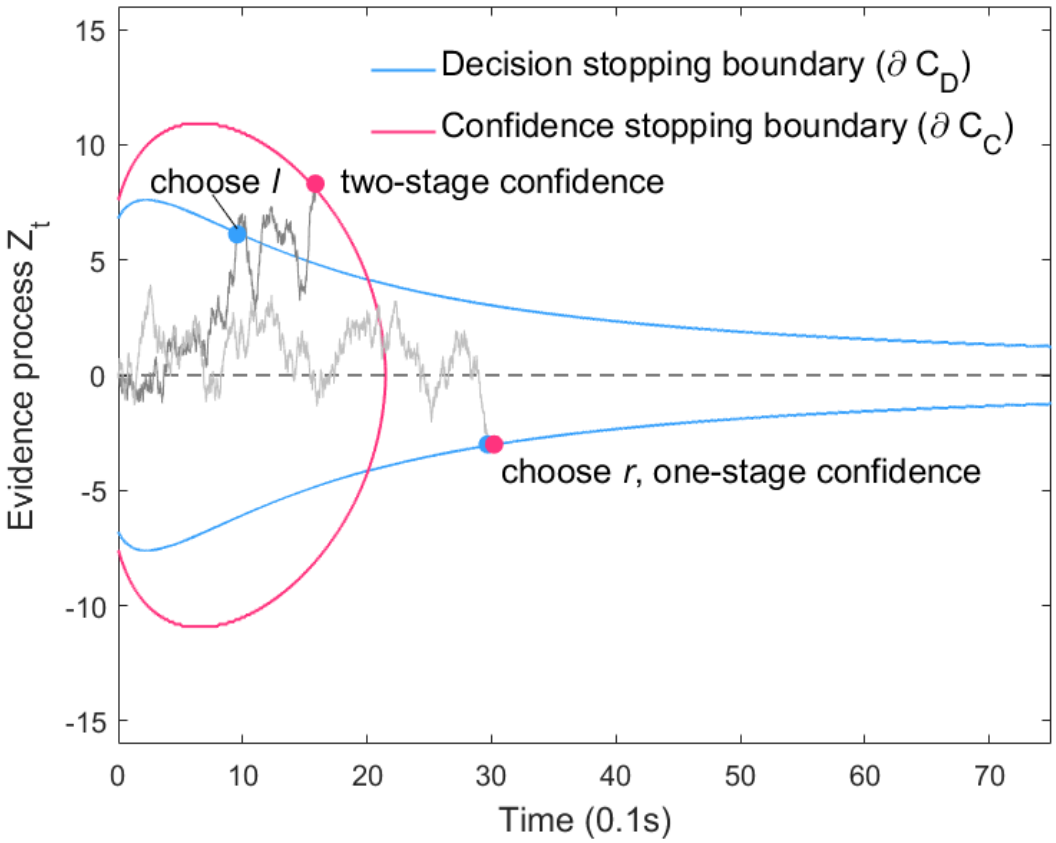


Figure 1.1: An illustration of optimal stopping of the evidence process for decision and confidence for a **low value of c̄**. The decision and confidence stopping boundaries are computed numerically (see Appendix 1.D) for parameters $\bar{c} = 0.007, c = 0.02, \alpha = 2, \sigma_0 = 1.8, X_0 = 0$. The grey lines represent two possible realizations of the evidence process. The blue dots represent the moments of decision formation and the pink dots represent the moments of confidence formation.

ing to zero at infinity and the confidence stopping boundary is a left-truncated ellipse.[19] These are the typical features of these regions (for another example of the boundaries with different parameters, see Figure 1.G.1 in Appendix 1.G). One less typical feature of the boundaries depicted in Figure 1.1 is that they first expand before collapsing.[20] We chose this non-typical (but realistic) set of parameters to illustrate an important point developed later in this section.

[19]For the characterization of the decision stopping boundary, see Fudenberg, Strack, and Strzalecki (2018). The shape of the confidence stopping boundary is driven by region $\mathring{C}_C$ derived analytically in Appendix 1.A.5.

[20]This can be seen from Figure 3 in the Online Appendix of Fudenberg, Strack, and Strzalecki (2018). The estimated decision boundaries for most participants are monotonically collapsing.

The value of $\bar{c}$ used in Figure 1.1 is sufficiently low so that it is sometimes optimal to sample beyond decisions. In particular, if a realization of the evidence process Z_t is sufficiently strong, e.g., as depicted by the dark grey line, then the decision is made inside the confidence continuation region C_C and it is optimal to sample beyond such decision to refine confidence. Hence, the confidence judgement will be based on more information than the decision. We call such cases of informational dissociation between decision and confidence *two-stage confidence*.

On the other hand, if a realization of the evidence process Z_t is sufficiently weak, e.g., as depicted by the light grey line, then the decision is made outside the confidence continuation region C_C and it is optimal to stop as soon as the decision is made. In fact, from the perspective of confidence, it would be optimal to stop even sooner in this case—upon hitting ∂C_C; however, the confidence stopping time is chosen only from a restricted set of stopping times that come after the optimal decision stopping time. Hence, the confidence judgement will be based on the same information as the decision. We call such cases of informational congruence between decision and confidence *one-stage confidence*.

As one might expect, the higher is $\bar{c}$, the smaller is the confidence continuation region C_C, thus the less likely confidence will be two-stage. Moreover, from the shape of the decision and confidence stopping boundaries, one might expect that the two-stage confidence would prevail only for the fastest decisions. However, as suggested by the numerical solution in the next paragraph, this intuition might not hold. This might be an important point speaking to the architecture underlying confidence formation that would be missed by heuristic models of decision confidence (Moran, Teodorescu, and Usher 2015; Pleskac and Busemeyer 2010).

In Figure 1.2, we depict the numerically computed decision and confidence stopping boundaries for an individual with $\bar{c} = 0.012$ and the same remaining parameters as in Figure 1.1, $c = 0.02, \alpha = 2, \sigma_0 = 1.8, X_0 = 0$. The value of $\bar{c}$ is sufficiently low so that there is space for two-stage confidence. However, for the fastest decisions, C_C is contained in C_D, so confidence will be one-stage for these decisions—for example, see the strong realization of the evidence process depicted in dark grey. On the other hand, the weak realization of the evidence process depicted in light grey will lead to a relatively fast (but not the fastest) decision and two-stage confidence.[21] Finally, the weakest realizations

[21]Notice that even though the decision and confidence boundaries ∂C_D and ∂C_C are not very distant

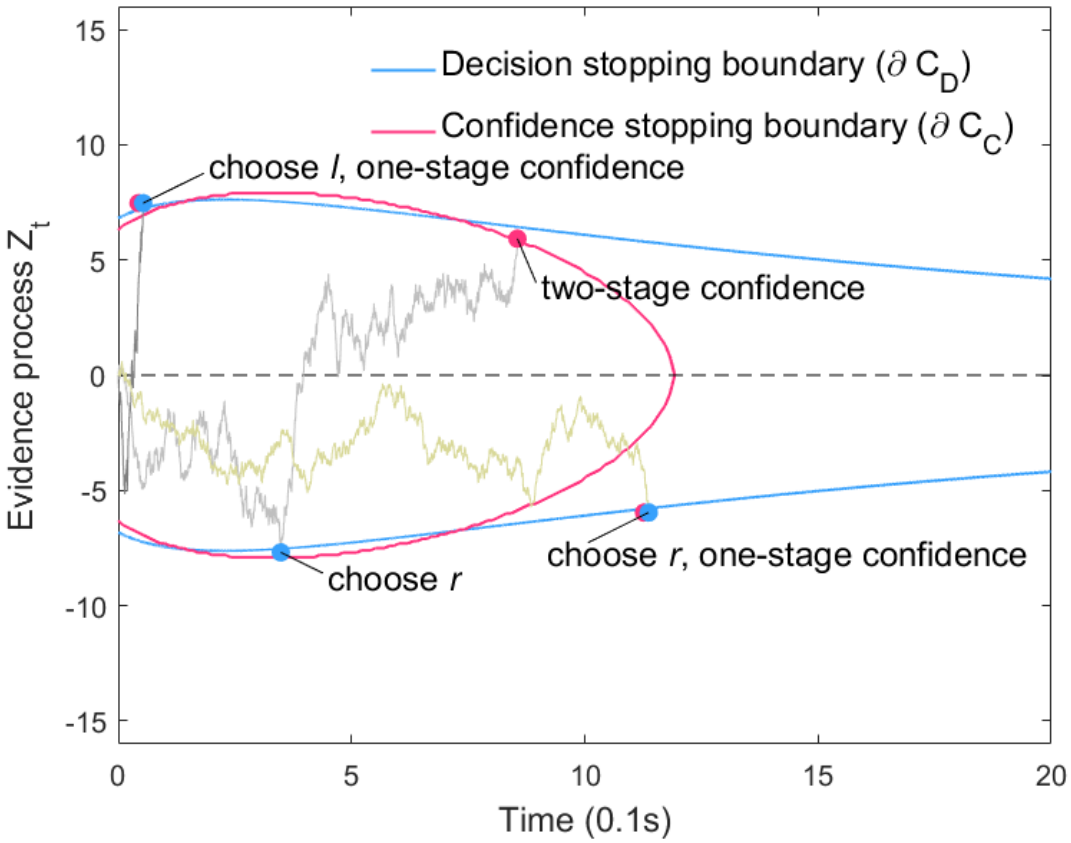


Figure 1.2: An illustration of optimal stopping of the evidence process for decision and confidence for a **moderately high value of c̄**. The decision and confidence stopping boundaries are computed numerically (see Appendix 1.D) for parameters $\bar{c} = 0.012, c = 0.02, \alpha = 2, \sigma_0 = 1.8, X_0 = 0$. The grey and yellow lines represent three possible realizations of the evidence process. The blue dots represent the moments of decision formation and the pink dots represent the moments of confidence formation.

of the evidence process (as the yellow one) will again lead to one-stage confidence. This illustrates a potential non-monotonicity of occurrence of two-stage confidence.

In summary, the numerical solutions suggest several qualitative insights about the informational underpinning of confidence judgements relative to decisions:

- the slowest decisions are associated with one-stage confidence, i.e., decisions and decision confidence are based on the same evidence;
- if the cost of time relative to the importance of good confidence $\bar{c}$ is sufficiently

in the region for intermediately fast decisions, the resulting timing of decision and confidence formation may be very distinct as illustrated by the light grey realization of the evidence process.

low, (intermediately) fast decisions are associated with two-stage confidence, i.e., decision confidence is based on more evidence than decisions;

- nevertheless, the fastest decisions might still be associated with one-stage confidence for intermediately low values of $\bar{c}$ under some parametrizations (namely, α, σ_0, and c).

The first claim is directly supported analytically by Theorem 1.1. The second claim follows from considering negligible confidence sampling cost ($\bar{c} \to 0$) while keeping c, α, and σ_0 fixed (at non-zero levels): in this case, post-decisional information acquisition becomes optimal after each decision. Hence, there exists a lower bound for $\bar{c}$, below which all fast decisions lead to post-decisional information acquisition. The third claim is a conjecture motivated by the numerical solutions. It seems to be difficult to establish analytically exact thresholds for "intermediate" values of $\bar{c}$ and characterize the situation for the fastest decisions. As an immediate corollary of Theorem 1.1, we can at least establish an upper bound for the threshold delineating "intermediate" values of $\bar{c}$: $\frac{\sigma_0^2}{2\pi\alpha^2}$ (a lower bound is trivially zero). From Theorem 1.1, we can see that if $\bar{c}$ is larger than $\frac{\sigma_0^2}{2\pi\alpha^2}$, there will be no scope for post-decisional information acquisition whatsoever.[22]

1.5 Discussion

Our analysis suggests that the occurrence of post-decisional information acquisition for confidence refinement may follow a more nuanced pattern than has been previously imagined in the cognitive science literature. Specifically, this literature has considered only two extreme theories speaking to post-decisional information acquisition: one-stage theories—with no post-decisional information acquisition—and two-stage theories—with post-decisional information acquisition after each decision. Our insights reconcile these

[22]Besides the numerical solution (see Appendix 1.D), we can also derive analytically the first-order condition for deterministic stopping, see Appendix 1.B, which is necessary but not sufficient. The interest of studying the deterministic stopping lies in the following implication: if immediate stopping is not optimal in the set of deterministic stopping times, then it is not optimal in the unconstrained set of all stopping times $\mathcal{T}$ either. Nevertheless, numerical solutions suggest that the reverse implication does not hold, so the analysis of deterministic stopping times likely provides only a subset of the full continuation region.

two competing views by elucidating how both modes of decision confidence formation may manifest themselves under one underlying mechanism.

In this section, we evaluate our model (referred to as Two-Stage Sequential Sampling or 2SS) on the basis of empirical patterns of Moran, Teodorescu, and Usher (2015) (which extend the list of patterns of Pleskac and Busemeyer [2010]). Moreover, we contrast the 2SS model with the most related models.

1.5.1 Empirical evaluation and contrast to heuristic models

Using a small simulation study, we report in Table 1.E.1 in Appendix 1.E that 2SS matches all empirical patterns of Moran, Teodorescu, and Usher (2015, Table 1, p. 102) except for the positive correlation of decision and interjudgement times (pattern 8.4). Intuitively, there is scope for non-zero interjudgement times in 2SS only for relatively fast decisions.[23]

The 2SS model can be seen as an attempt to microfound the heuristic model of Moran, Teodorescu, and Usher (2015) (referred to as Collapsing Confidence Boundary or CCB). CCB matches all patterns in Table 1.E.1. In particular, it captures the positive correlation of decision and interjudgement times, unlike 2SS. The driving force behind this feature is non-stationarity of the confidence boundary and selection: in CCB, the confidence boundary is decision-triggered, i.e., it is placed only at the moment of decision. Hence, the distance the evidence particle has to travel to reach a given point on the confidence boundary is independent of the decision stopping point (unlike in the 2SS model). Therefore, the selection effect operates: slow decisions are more likely to be generated by a signal with low drift (in absolute value) and this carries over to the confidence stage, leading to slower confidence judgements more likely as well.

The decision-triggered confidence boundary in CCB reflects the assumption that confidence is always two-stage. In contrast, in this paper we do not impose this assumption; instead, we allow for the second stage and derive when it should actually occur. Thus, our paper is an attempt to give a statistical foundation for the two-stage confidence architec-

[23] 65% and 49% trials lead to two-stage confidence (non-zero interjudgement times) in 2SS under the two parametrizations used in the simulations specified in Table 1.E.1. Hence, there is a non-trivial use of the two-stage-confidence mode.

ture. Our analysis reveals that this uniform assumption may not be well justified and a strict distinction between one-stage and two-stage theories of decision confidence (Moran, Teodorescu, and Usher 2015) might be misleading. Both modes of confidence formation may manifest themselves under one underlying mechanism. Moreover, they may manifest themselves in a surprising manner as in Figure 1.2, i.e., one-stage confidence for the fastest and slow decisions while two-stage for intermediately fast ones.[24]

It is worth noting that even the empirical results of Moran, Teodorescu, and Usher (2015) "may indicate that there are important individual differences with respect to the number of information collection stages" (p. 111). Our approach enables us to study more rigorously not only these individual differences (e.g., different values of mental noise α [Enke and Graeber, 2022]), but also differences for a given individual across decision environments (e.g., different opportunity cost c), and even differences in the use of one-stage vs. two-stage mode for a given individual in a given decision environment (for fixed parameters).

1.5.2 Comparison with the model of Fudenberg, Strack, and Strzalecki (2018)

Since the 2SS model is an extension of the uncertain-difference model of Fudenberg, Strack, and Strzalecki (2018) (referred to as FSS), it is natural to discuss predictions of FSS about decision confidence. As outlined in the Introduction, the natural way to define decision confidence in the FSS model is by (1.9) at the time of decision τ. However, such model is incompatible with the observed ability of people to recognize their own errors even without explicit feedback (Yeung and Summerfield 2012; Fleming and Daw 2017). Moreover, Moran, Teodorescu, and Usher (2015) conduct experiments where post-decisional evidence availability has a causal effect on confidence resolution.[25] These experiments put forward the interjudgement time (RT2) as an important variable

[24] Moran, Teodorescu, and Usher (2015) build on the work of Pleskac and Busemeyer (2010) who also propose heuristic models of decision confidence capturing the two-stage idea. Importantly, these insights provided by 2SS in relation to CCB apply to the models of Pleskac and Busemeyer (2010) too.

[25] Confidence resolution is the correlation between confidence and decision correctness.

to guide the modeling of decision confidence—a variable that FSS is silent about.[26]

In Table 1.E.1, we report that FSS does not capture the increased confidence resolution under time pressure (higher c). Intuitively, decisions and confidence are tied in FSS, so time pressure affects both negatively. On the other hand, in two-stage models, time pressure makes the job easier for the confidence system because mistakes are easier to recognize.

1.5.3 Simpler version: two-stage Wald model

One may wonder whether our extension to the model of Fudenberg, Strack, and Strzalecki (2018) would bring something interesting in the simpler Wald model. We develop this idea in Appendix 1.C; call it *two-stage Wald model*. We find that the decision maker either (i) never wants to sample beyond decisions (one-stage confidence always) and there is a single confidence level achieved, or (ii) always wants to sample beyond decisions (two-stage confidence always) and there are two possible confidence levels achieved—one above $\frac{1}{2}$ for correctness confirmation and one below $\frac{1}{2}$ for error recognition.[27] The use of mode (i) or (ii) depends on the parameters. In particular, post-decisional information acquisition (mode (ii)) is favored by higher c and lower $\bar{c}$.

People are likely to report more than two levels of confidence, so to account for that we can introduce a noisy readout of confidence, i.e., the confidence levels predicted by the two-stage Wald model would be reported with some independent noise. Nevertheless, there are several empirical regularities that such model cannot explain. For example, there will be no correlation between decision times and confidence because the belief process is a time-homogeneous diffusion. Intuitively, conditional on stopping for decision, the belief process restarts at the same point regardless of the decision time due to the constant decision boundary. Independent readout noise does not change this relationship. This is inconsistent with the empirical regularity no. 4 from Moran, Teodorescu, and Usher (2015) (negative correlation between decision time and confidence). Moreover, the regularities tied to the manipulation of stimulus discriminability will pose problems

[26]The inability to capture error monitoring and the effect of post-decisional evidence availability is shared by all single-stage models, e.g., the Wald model.

[27]In fact, these predictions would be obtained with an even simpler static model of a rationally inattentive decision maker. However, the static version would be silent about the reaction times.

for the two-stage Wald model too, e.g., pattern no. 3 of negative correlation between confidence and difficulty.[28]

1.6 Conclusion

We develop a normative dynamic model of decision confidence. This model captures the dual-system view of mind with the decision system aiming to maximize expected utility and the confidence system aiming to monitor the decision system and provide good feedback on the decision system's performance. We focus on studying optimal informational dissociations between decisions and decision confidence, i.e., we ask when we should expect the decision and confidence to be based on the same evidence and when not. We question the following assumption adopted by models from cognitive science: decision confidence is always (i.e., uniformly across all decisions) based on post-decisional evidence accumulation. We find that this assumption is not justified by our normative model; our model suggests that there is scope for post-decisional evidence accumulation only for relatively fast decisions. Moreover, a nontrivial pattern may emerge in some situations—confidence based on the same information as decisions for very fast and slow decisions and post-decisional evidence accumulation for intermediately fast decisions.

Our findings contribute to the theory of metacognitive monitoring. In particular, we provide a normative foundation for post-decisional evidence accumulation and derive how one- and two-stage modes of decision confidence may arise under a single mechanism. Moreover, an empirical evaluation of our model indicates that it is consistent with the patterns of Moran, Teodorescu, and Usher (2015), except for the positive correlation of decision times and interjudgement times. This and similar patterns may guide future theorizing about the mechanisms underlying decision confidence, which will lead to a better understanding of its functional role. Moreover, it will be interesting to contrast and merge our approach to theorizing about mechanisms behind decision confidence with the complementary direction of "second-order" models (Fleming and Daw 2017), in which

[28] Nevertheless, one has to be cautious about the use of these empirical regularities because manipulating the objective difference between the two options across trials does not fit the setup of the Wald model. In the Wald model, the decision maker is assumed to know the difference between the options a priori; she is uncertain only about which option is the better one. Hence, the prior of the decision maker is not correct in this setup and there is a mismatch between the experimenter and decision maker.

distinct but related processes inform decision and decision confidence. Finally, decision confidence is involved in the control of future mental processes and behavior. Different mechanisms underlying decision confidence may lead to different predictions about the relationships between confidence and future behavior (Schulz, Fleming, and Dayan 2021). Hence, analyzing the implications of our model of decision confidence for the control of future behavior is an exciting avenue for future research.

1.A Technical details and proofs

1.A.1 Decision objective

We aim to clarify the objective of the decision system that appeared in various forms in the literature. Let us recollect our formulation of the decision problem (1.3):

$$\sup_{\tau \in \mathcal{T}} \mathrm{E}\left[|\theta| \mathbb{1}\{\mathrm{sgn}(X_\tau) = \mathrm{sgn}(\theta)\} - c\tau\right].$$

Chernoff (1961) [see his equation (3.5)] and Zhitlukhin and Muravlev (2013) [see their p. 708] study the problem of minimization of the regret functional

$$\inf_{\tau \in \mathcal{T}} \mathrm{E}\left[k|\theta|\varepsilon(\theta) + \tilde{c}\tau\right],$$

where k is a constant and $\varepsilon(\theta)$ is the probability of error. We can write

$$\begin{aligned}
\mathrm{E}\left[k|\theta|\varepsilon(\theta) + \tilde{c}\tau\right] &= \mathrm{E}\left[k|\theta|\,\mathsf{P}\left(\mathrm{sgn}(X_\tau) \neq \mathrm{sgn}(\theta)|\theta\right) + \tilde{c}\tau\right] \\
&= \mathrm{E}\left[k|\theta|\mathrm{E}\left[\mathbb{1}\{\mathrm{sgn}(X_\tau) \neq \mathrm{sgn}(\theta)\}|\theta\right] + \tilde{c}\tau\right] \\
&= \mathrm{E}\left[k|\theta|(1 - \mathrm{E}\left[\mathbb{1}\{\mathrm{sgn}(X_\tau) = \mathrm{sgn}(\theta)\}|\theta\right]) + \tilde{c}\tau\right] \\
&= \mathrm{E}\left[k|\theta|\right] - \mathrm{E}\left[k|\theta|\mathrm{E}\left[\mathbb{1}\{\mathrm{sgn}(X_\tau) = \mathrm{sgn}(\theta)\}|\theta\right] - \tilde{c}\tau\right] \\
&= \mathrm{E}\left[k|\theta|\right] - \mathrm{E}\left[k\mathrm{E}\left[|\theta|\mathbb{1}\{\mathrm{sgn}(X_\tau) = \mathrm{sgn}(\theta)\}|\theta\right] - \tilde{c}\tau\right] \\
&= \mathrm{E}\left[k|\theta|\right] - \mathrm{E}\left[k|\theta|\mathbb{1}\{\mathrm{sgn}(X_\tau) = \mathrm{sgn}(\theta)\} - \tilde{c}\tau\right] \\
&= \mathrm{E}\left[k|\theta|\right] - k\mathrm{E}\left[|\theta|\mathbb{1}\{\mathrm{sgn}(X_\tau) = \mathrm{sgn}(\theta)\} - c\tau\right], \qquad (1.13)
\end{aligned}$$

where $c = \frac{\tilde{c}}{k}$ and we used the Law of Iterated Expectations in the second to last equality. Hence, the two problems lead to the same optimal stopping times.

On the other hand, Fudenberg, Strack, and Strzalecki (2018) [see their equation (6)] formulate their problem as utility maximization

$$\sup_{\tau \in \mathcal{T}} \mathrm{E}\left[\max\{X_\tau^l, X_\tau^r\} - c\tau\right],$$

where $X_t^i = \mathrm{E}\left[\theta^i|\mathcal{F}_t\right]$ is the posterior expectation of the utility of option $i \in \{l, r\}$. In

their Proposition 2, they show that

$$\begin{aligned}\mathrm{E}\left[\max\{X_\tau^l, X_\tau^r\} - c\tau\right] = \mathrm{E}\left[-\mathbb{1}\{X_\tau^l \geq X_\tau^r\}(\theta^r - \theta^l)^+ - \mathbb{1}\{X_\tau^r > X_\tau^l\}(\theta^l - \theta^r)^+ - c\tau\right] \\ + \mathrm{E}\left[\max\{\theta^l, \theta^r\}\right], \end{aligned} \tag{1.14}$$

where x^+ denotes the positive part of x, i.e., $x^+ = \max\{x, 0\}$. By denoting $\theta = \theta^l - \theta^r$ and $X_t = X_t^l - X_t^r = \mathrm{E}\left[\theta|\mathcal{F}_t\right]$, we can rewrite

$$-\mathbb{1}\{X_\tau^l \geq X_\tau^r\}(\theta^r - \theta^l)^+ - \mathbb{1}\{X_\tau^r > X_\tau^l\}(\theta^l - \theta^r)^+ = -\mathbb{1}\{X_\tau \geq 0\}(-\theta)^+ - \mathbb{1}\{X_\tau < 0\}\theta^+.$$

By adding $|\theta| = \mathbb{1}\{X_\tau \geq 0\}|\theta| + \mathbb{1}\{X_\tau < 0\}|\theta|$ to this expression, we obtain

$$\mathbb{1}\{X_\tau \geq 0\}(|\theta| - (-\theta)^+) + \mathbb{1}\{X_\tau < 0\}(|\theta| - \theta^+).$$

Since $|\theta| = \theta^+ + (-\theta)^+$, we get from (1.14)

$$\begin{aligned}&\mathrm{E}\left[\max\{X_\tau^l, X_\tau^r\} - c\tau\right] - \mathrm{E}\left[\max\{\theta^l, \theta^r\}\right] + \mathrm{E}\left[|\theta^l - \theta^r|\right] \\ &\quad = \mathrm{E}\left[\mathbb{1}\{X_\tau \geq 0\}\theta^+ + \mathbb{1}\{X_\tau < 0\}(-\theta)^+ - c\tau\right] \\ &\quad = \mathrm{E}\left[\mathbb{1}\{X_\tau \geq 0\}\mathbb{1}\{\theta \geq 0\}|\theta| + \mathbb{1}\{X_\tau < 0\}\mathbb{1}\{\theta < 0\}|\theta| - c\tau\right] \\ &\quad = \mathrm{E}\left[|\theta|\mathbb{1}\{\mathrm{sgn}(X_\tau) = \mathrm{sgn}(\theta)\} - c\tau\right],\end{aligned}$$

where we put $\mathrm{sgn}(0) = 1$ to take care of the case $X_\tau = 0$. Therefore, we finally obtain

$$\mathrm{E}\left[\max\{X_\tau^l, X_\tau^r\} - c\tau\right] - \mathrm{E}\left[\min\{\theta^l, \theta^r\}\right] = \mathrm{E}\left[|\theta|\mathbb{1}\{\mathrm{sgn}(X_\tau) = \mathrm{sgn}(\theta)\} - c\tau\right] \tag{1.15}$$

and we see that the problem of Fudenberg, Strack, and Strzalecki (2018) leads to the same optimal stopping times as our problem as long as the learning in their problem is only about the difference θ.

From equations (1.13), (1.14), and (1.15), we can see that the various objectives differ only in the baseline. The objective of Fudenberg, Strack, and Strzalecki is formulated as utility maximization, so the baseline is zero. Chernoff's objective is formulated as regret minimization, so the baseline is the maximum attainable utility, which can be best seen from (1.14). Our objective is formulated as maximization of correctness and the baseline is the minimum attainable utility, which can be best seen from (1.15).

1.A.2 Beliefs

Lemma 1.A.1. *Let $\theta \sim N(X_0, 2\sigma_0^2)$ and $Z_t = \theta t + \alpha\sqrt{2}B_t$, $t \geq 0$, where $\alpha > 0$ is a parameter and B_t is a standard Brownian motion independent of θ. Then the posterior distribution of θ after observing $Z_s, s \leq t$, is normal with mean*

$$\frac{\sigma_0^{-2}X_0 + \alpha^{-2}Z_t}{\sigma_0^{-2} + \alpha^{-2}t} \tag{1.16}$$

and variance

$$\frac{2}{\sigma_0^{-2} + \alpha^{-2}t}. \tag{1.17}$$

Proof. As outlined by Chernoff (1961, p. 81) (and developed in more detail by Shiryaev[29]), the conditional distribution of θ is determined by

$$\mathsf{P}\left(\theta \leq y | \mathcal{F}_t\right) = \frac{\int_{-\infty}^{y} \frac{\mathrm{d}\,\mathsf{P}\left(Z_0^t|\theta=\xi\right)}{\mathrm{d}\,\mathsf{P}\left(Z_0^t|\theta=0\right)} \mathrm{dP}_\theta(\xi)}{\int_{-\infty}^{\infty} \frac{\mathrm{d}\,\mathsf{P}\left(Z_0^t|\theta=\xi\right)}{\mathrm{d}\,\mathsf{P}\left(Z_0^t|\theta=0\right)} \mathrm{dP}_\theta(\xi)},$$

where

$$\frac{\mathrm{d}\,\mathsf{P}\left(Z_0^t|\theta=\xi\right)}{\mathrm{d}\,\mathsf{P}\left(Z_0^t|\theta=0\right)}$$

is the Radon-Nikodym derivative of the measure of the process $Z_0^t = \{Z_s\}_{s=0}^t$ with $\theta = \xi$ with respect to the measure of the process $Z_0^t = \{Z_s\}_{s=0}^t$ with $\theta = 0$. The Radon-Nikodym derivative can be calculated explicitly as[30]

$$\frac{\mathrm{d}\,\mathsf{P}\left(Z_0^t|\theta=\xi\right)}{\mathrm{d}\,\mathsf{P}\left(Z_0^t|\theta=0\right)} = e^{\frac{1}{2}\left(\frac{\xi}{\alpha^2}Z_t - \frac{1}{2}\frac{\xi^2}{\alpha^2}t\right)}.$$

Hence, the conditional density of θ is

$$p(y; t, Z_t) = \frac{\mathrm{d}\,\mathsf{P}\left(\theta \leq y|\mathcal{F}_t\right)}{\mathrm{d}y} = \frac{e^{\frac{1}{2}\left(\frac{y}{\alpha^2}Z_t - \frac{1}{2}\frac{y^2}{\alpha^2}t\right)}p(y)}{\int_{-\infty}^{\infty} e^{\frac{1}{2}\left(\frac{\xi}{\alpha^2}Z_t - \frac{1}{2}\frac{\xi^2}{\alpha^2}t\right)}p(\xi)\,\mathrm{d}\xi}, \tag{1.18}$$

[29] See p. 8–9 at https://www.uni-ulm.de/fileadmin/website_uni_ulm/mawi.inst.110/lehre/ws13/Workshop_Probab_Anal_Geom/Shiryaev.pdf.

[30] For example, see the Girsanov theorem, specifically Theorem 8.6.4 in Øksendal (2003).

where p is the prior density of θ. The numerator is

$$\begin{aligned}
&\frac{1}{\sqrt{4\pi\sigma_0^2}} e^{\frac{1}{2}\left(\frac{y}{\alpha^2}Z_t - \frac{1}{2}\frac{y^2}{\alpha^2}t - \frac{1}{2}\frac{(y-X_0)^2}{\sigma_0^2}\right)} \\
&= \frac{1}{\sqrt{4\pi\sigma_0^2}} e^{-\frac{1}{4\sigma_0^2}(1+\sigma_0^2\alpha^{-2}t)\left(y^2 - 2y\frac{\sigma_0^2\alpha^{-2}Z_t+X_0}{1+\sigma_0^2\alpha^{-2}t} + \frac{X_0^2}{1+\sigma_0^2\alpha^{-2}t} + \left[\left(\frac{\sigma_0^2\alpha^{-2}Z_t+X_0}{1+\sigma_0^2\alpha^{-2}t}\right)^2 - \left(\frac{\sigma_0^2\alpha^{-2}Z_t+X_0}{1+\sigma_0^2\alpha^{-2}t}\right)^2\right]\right)} \\
&= \frac{1}{\sqrt{4\pi\sigma_0^2}} e^{-\frac{1}{4\sigma_0^2}(1+\sigma_0^2\alpha^{-2}t)\left(y - \frac{\sigma_0^2\alpha^{-2}Z_t+X_0}{1+\sigma_0^2\alpha^{-2}t}\right)^2} e^{\frac{-(1+\sigma_0^2\alpha^{-2}t)}{4\sigma_0^2}\left[\frac{X_0^2}{1+\sigma_0^2\alpha^{-2}t} - \left(\frac{\sigma_0^2\alpha^{-2}Z_t+X_0}{1+\sigma_0^2\alpha^{-2}t}\right)^2\right]}. \qquad (1.19)
\end{aligned}$$

Therefore, by exploiting the density of normal distribution with mean

$$\frac{X_0 + \sigma_0^2\alpha^{-2}Z_t}{1+\sigma_0^2\alpha^{-2}t}$$

and variance

$$\frac{2\sigma_0^2}{1+\sigma_0^2\alpha^{-2}t},$$

the denominator in (1.18) is

$$e^{\frac{-(1+\sigma_0^2\alpha^{-2}t)}{4\sigma_0^2}\left[\frac{X_0^2}{1+\sigma_0^2\alpha^{-2}t} - \left(\frac{\sigma_0^2\alpha^{-2}Z_t+X_0}{1+\sigma_0^2\alpha^{-2}t}\right)^2\right]} \frac{1}{\sqrt{1+\sigma_0^2\alpha^{-2}t}}. \qquad (1.20)$$

Hence, by putting (1.18), (1.19), and (1.20) together, we can see that the conditional density is

$$p(y;t,Z_t) = \frac{1}{\sqrt{2\pi\frac{2\sigma_0^2}{1+\sigma_0^2\alpha^{-2}t}}} e^{-\frac{1}{4\sigma_0^2}(1+\sigma_0^2\alpha^{-2}t)\left(y - \frac{\sigma_0^2\alpha^{-2}Z_t+X_0}{1+\sigma_0^2\alpha^{-2}t}\right)^2},$$

which is the density of the normal distribution with mean

$$\frac{X_0 + \sigma_0^2\alpha^{-2}Z_t}{1+\sigma_0^2\alpha^{-2}t} = \frac{\sigma_0^{-2}X_0 + \alpha^{-2}Z_t}{\sigma_0^{-2}+\alpha^{-2}t}$$

and variance

$$\frac{2\sigma_0^2}{1+\sigma_0^2\alpha^{-2}t} = \frac{2}{\sigma_0^{-2}+\alpha^{-2}t}.$$

□

1.A.3 Structure of the unrestricted confidence stopping problem

Let us recall the full unrestricted confidence stopping problem (1.12)

$$\inf_{\tau' \in \mathcal{T}} \mathrm{E}\left[\Phi\left(\frac{\sigma_0^{-2}X_0 + \alpha^{-2}Z_{\tau'}}{\sqrt{2(\sigma_0^{-2} + \alpha^{-2}\tau')}}\right)\Phi\left(-\frac{\sigma_0^{-2}X_0 + \alpha^{-2}Z_{\tau'}}{\sqrt{2(\sigma_0^{-2} + \alpha^{-2}\tau')}}\right) + \bar{c}\tau'\right].$$

Here, we discuss the structure of this problem.

We can restart process $\{Z_t\}_{t\geq 0}$ after time s

$$Z_{s+t} = \theta(s+t) + \alpha\sqrt{2}B_{s+t} = Z_s + \theta t + \alpha\sqrt{2}(B_{s+t} - B_s),$$

where we denote the new starting point $z := Z_s$ and the new standard Brownian motion $\widetilde{B}_t := B_{s+t} - B_s$ for $t \geq 0$. This motivates us to introduce process $\{Z_t^z\}_{t\geq 0}$ with the same differential as $\{Z_t\}_{t\geq 0}$, but starting at z

$$Z_t^z = z + \theta t + \alpha\sqrt{2}\widetilde{B}_t,\ t \geq 0.$$

The innovation representation (Liptser and Shiryaev 2000, Section 7.4) of $\{Z_t^z\}_{t\geq 0}$ starts with a different initial information set, $\widetilde{\mathcal{F}}_t = \mathcal{F}_{s+t}$ for $t \geq 0$, thus the prior is $N(X_s, \sigma_s^2)$. Hence, its innovation representation is

$$Z_t^z = z + \int_0^t \widetilde{X}_r\,\mathrm{d}r + \alpha\sqrt{2}\overline{\widetilde{B}}_t,$$

where

$$\widetilde{X}_r = \mathrm{E}\left[\theta|\widetilde{\mathcal{F}}_r\right] = \frac{\sigma_0^{-2}X_0 + \alpha^{-2}Z_r^z}{\sigma_0^{-2} + \alpha^{-2}(s+r)}$$

and

$$\overline{\widetilde{B}}_t = \frac{1}{\alpha\sqrt{2}}\left(\theta t + \alpha\sqrt{2}\widetilde{B}_t - \int_0^t \widetilde{X}_r\,\mathrm{d}r\right)$$

is a standard Brownian motion.

We can introduce process $\{Y_t^{(s,z)}\}_{t\geq 0}$ such that $Y_t^{(s,z)} = (s+t, Z_t^z)'$ for $t \geq 0$ with the differential

$$\mathrm{d}Y_t^{(s,z)} = \begin{pmatrix} 1 \\ \frac{\sigma_0^{-2}X_0 + \alpha^{-2}Z_t^z}{\sigma_0^{-2} + \alpha^{-2}(s+t)} \end{pmatrix}\mathrm{d}t + \begin{pmatrix} 0 \\ \alpha\sqrt{2} \end{pmatrix}\mathrm{d}\overline{\widetilde{B}}_t,\ t \geq 0,\ Y_0^{(s,z)} = (s,z). \tag{1.21}$$

Since this process is Markovian (it always restarts anew), the unrestricted confidence stopping problem (1.12) has the structure of the problem (2.2.2) of Peskir and Shiryaev (2006). However, their condition (2.2.1) is not satisfied in our problem because of the potentially unbounded sampling cost $\bar{c}t$. Nevertheless, their comments on p. 27 and 2 indicate that this may not be a problem, especially if we restrict our search for optimal stopping times to $\tau' \in \mathcal{T}$ such that $\mathrm{E}[\tau'] < \infty$ for which the expectation of the cost upon stopping is well defined. Restriction to such stopping times is innocuous in our problem (1.12) because, by boundedness of the function $x \mapsto \Phi(x)\Phi(-x)$, stopping times that are expected to be infinite are dominated by immediate stopping due to unbounded sampling cost.

1.A.4 Notation

To simplify notation further, we introduce the loss function $f\colon [0,\infty) \times \mathbb{R} \to \mathbb{R}$ defined for all $(r, x) \in [0,\infty) \times \mathbb{R}$ as follows

$$f(r,x) = \Phi\left(\frac{\sigma_0^{-2}X_0 + \alpha^{-2}x}{\sqrt{2(\sigma_0^{-2} + \alpha^{-2}r)}}\right)\Phi\left(-\frac{\sigma_0^{-2}X_0 + \alpha^{-2}x}{\sqrt{2(\sigma_0^{-2} + \alpha^{-2}r)}}\right) + \bar{c}r. \tag{1.22}$$

Further, we denote the expected loss at time $t \in [0,\infty)$ when we start sampling with $s \in [0,\infty)$ amount of evidence and the evidence process at position $z \in \mathbb{R}$

$$u(t;s,z) = \mathrm{E}\left[f(s+t, Z_t^z)\right], \tag{1.23}$$

where process Z^z is introduced in Section 1.A.3. Finally, we denote the value function

$$U(s,z) = \inf_{\tau\in\mathcal{T}} u(\tau;s,z). \tag{1.24}$$

With this notation, the unconstrained confidence stopping problem can be formulated as the problem of characterizing the set

$$C_C = \{(s,z) \in [0,\infty) \times \mathbb{R} : U(s,z) < f(s,z)\}.$$

1.A.5 Subcontinuation region for the unrestricted confidence stopping problem

Lemma 1.A.2. *Let*

$$\mathring{C}_C = \left\{(r,x) \in [0,\infty) \times \mathbb{R} \colon \frac{\alpha^{-2}}{\sigma_0^{-2} + \alpha^{-2} r} \varphi^2 \left(\frac{\sigma_0^{-2} X_0 + \alpha^{-2} x}{\sqrt{2(\sigma_0^{-2} + \alpha^{-2} r)}} \right) > \bar{c} \right\}. \tag{1.25}$$

It is not optimal to stop sampling for confidence in this region, i.e., $\mathring{C}_C \subseteq C_C$.

Proof. Consider process $\{Y_t^{(s,z)}\}_{t \geq 0}$ introduced in (1.21) and function f introduced in (1.22). By applying the Itô formula, we obtain

$$f(Y_t^{(s,z)}) = f(s,z) + \int_0^t A f(s+r, Z_r^z) \, \mathrm{d}r + \int_0^t \alpha \sqrt{2} \frac{\partial f}{\partial x}(s+r, Z_r^z) \, \mathrm{d}\overline{\overline{B}}_r, \tag{1.26}$$

where A is a differential operator acting on f such that, when evaluated at (r,x),

$$A f(r,x) = \frac{\partial f}{\partial r}(r,x) + \frac{\sigma_0^{-2} X_0 + \alpha^{-2} x}{\sigma_0^{-2} + \alpha^{-2} r} \frac{\partial f}{\partial x}(r,x) + \alpha^2 \frac{\partial^2 f}{\partial x^2}(r,x). \tag{1.27}$$

Similar to Lemma 7.3.2 of Øksendal (2003), it can be proved that for a stopping time τ' such that $\mathrm{E}[\tau'] < \infty$,

$$\mathrm{E}\left[\int_0^{\tau'} \alpha \sqrt{2} \frac{\partial f}{\partial x}(s+r, Z_r^z) \, \mathrm{d}\overline{\overline{B}}_r \right] = 0,$$

thus, from (1.26),

$$\mathrm{E}\left[f(Y_{\tau'}^{(s,z)}) \right] = f(s,z) + \mathrm{E}\left[\int_0^{\tau'} A f(s+r, Z_r^z) \, \mathrm{d}r \right]. \tag{1.28}$$

Consider set

$$\mathring{C}_C = \{(r,x) \in [0,\infty) \times \mathbb{R} \colon A f(r,x) < 0\}.$$

For $(s,z) \in \mathring{C}_C$ and a bounded open set V such that $(s,z) \in V \subset \mathring{C}_C$, consider stopping time

$$\tau_V = \inf\{t \geq 0 \colon Y_t^{(s,z)} \notin V\}.$$

Then, we can see from (1.28) that

$$\mathrm{E}\left[f(Y_{\tau_V}^{(s,z)})\right] < f(s,z).$$

Hence, as long as we are in $\mathring{C}_C$, it is not optimal to stop sampling for confidence because we expect the total confidence cost f to decrease at least until we exit from $\mathring{C}_C$. Therefore, $\mathring{C}_C \subseteq C_C$ and that is why we call $\mathring{C}_C$ a *subcontinuation region* for the unrestricted confidence stopping problem (1.12).[31]

To express region $\mathring{C}_C$ explicitly, denote

$$M(r,x) = \frac{\sigma_0^{-2}X_0 + \alpha^{-2}x}{\sqrt{2(\sigma_0^{-2} + \alpha^{-2}r)}}. \tag{1.29}$$

First, we calculate

$$M_r(r,x) = \frac{\partial M(r,x)}{\partial r} = -\frac{\alpha^{-2}}{2(\sigma_0^{-2} + \alpha^{-2}r)}M(r,x), \tag{1.30}$$

$$M_x(r,x) = \frac{\partial M(r,x)}{\partial x} = \frac{\alpha^{-2}}{\sqrt{2(\sigma_0^{-2} + \alpha^{-2}r)}}, \tag{1.31}$$

$$M_{xx}(r,x) = \frac{\partial^2 M(r,x)}{\partial x^2} = 0. \tag{1.32}$$

Second, using (1.32) and the following properties of the standard normal pdf φ

$$\varphi'(x) = -x\varphi(x),$$
$$\varphi(-x) = \varphi(x),$$

we obtain

$$\frac{\partial f(r,x)}{\partial r} = \varphi(M(r,x))M_r(r,x)(1 - 2\Phi(M(r,x))) + \bar{c}, \tag{1.33}$$

$$\frac{\partial f(r,x)}{\partial x} = \varphi(M(r,x))M_x(r,x)(1 - 2\Phi(M(r,x))), \tag{1.34}$$

$$\frac{\partial^2 f(r,x)}{\partial x^2} = \varphi(M(r,x))M_x^2(r,x)[-M(r,x)(1 - 2\Phi(M(r,x))) - 2\varphi(M(r,x))]. \tag{1.35}$$

Finally, plugging expressions (1.30)–(1.31) and (1.33)–(1.35) into (1.27) and simplifying

[31] In fact, $\mathring{C}_C \neq C_C$ in general. As Øksendal (2003, p. 205) writes, this is "the typical situation" in these problems.

yields

$$Af(r,x) = \bar{c} - \frac{\alpha^{-2}}{\sigma_0^{-2} + \alpha^{-2} r} \varphi^2(M(r,x)). \tag{1.36}$$

Hence, the subcontinuation region is

$$\mathring{C}_C = \left\{ (r,x) \in [0,\infty) \times \mathbb{R} \colon \frac{\alpha^{-2}}{\sigma_0^{-2} + \alpha^{-2} r} \varphi^2 \left(\frac{\sigma_0^{-2} X_0 + \alpha^{-2} x}{\sqrt{2(\sigma_0^{-2} + \alpha^{-2} r)}} \right) > \bar{c} \right\}.$$

□

1.A.6 Boundedness of the unrestricted confidence stopping time

Lemma 1.A.3. *Any optimal stopping time in the unconstrained confidence stopping problem* (1.12) *is bounded almost surely by* $\max\{\alpha^2[(2\pi\alpha^2\bar{c})^{-1} - \sigma_0^{-2}], 0\}$.

Proof. Intuitively, the bound is the rightmost point of $\mathring{C}_C$. Denote $T_c := \alpha^2[(2\pi\alpha^2\bar{c})^{-1} - \sigma_0^{-2}]$ and suppose $T_c > 0$.[32] Toward contradiction, suppose we have an optimal stopping time of the unconstrained confidence stopping problem (1.12), σ, that ends beyond T_c with strictly positive probability, i.e., $\mathrm{P}(\sigma > T_c) > 0$. Nevertheless, we still assume that $\mathrm{E}[\sigma] < \infty$ because stopping times that are expected to be infinite cannot be optimal as we already argued at the end of Section 1.A.3.[33]

With the use of notation from Section 1.A.5 and by using (1.28), the expected loss in problem (1.12) for σ can be written as

$$\mathrm{E}\left[f(Y_\sigma^{(0,0)})\right] = f(0,0) + \mathrm{E}\left[\int_0^\sigma Af(r, Z_r)\,\mathrm{d}r\right].$$

Similarly, we can write for stopping time $\sigma \wedge T_c$

$$\mathrm{E}\left[f(Y_{\sigma\wedge T_c}^{(0,0)})\right] = f(0,0) + \mathrm{E}\left[\int_0^{\sigma\wedge T_c} Af(r, Z_r)\,\mathrm{d}r\right].$$

[32] T_c is the (unique) solution of equation $Af(r, -\alpha^2\sigma_0^{-2}X_0) = 0$, where $Af(r,x)$ is given by (1.36) (the motivation for this choice will become obvious in the proof). Notice that we are looking for the solution $r \in \mathbb{R}$, so it can also be negative.

[33] As a consequence, we also have $\mathrm{P}(\sigma < \infty) = 1$.

Hence,

$$\mathrm{E}\left[f(Y_{\sigma}^{(0,0)})\right] - \mathrm{E}\left[f(Y_{\sigma\wedge T_c}^{(0,0)})\right] = \mathrm{E}\left[\int_0^{\sigma} Af(r, Z_r)\,\mathrm{d}r - \int_0^{\sigma\wedge T_c} Af(r, Z_r)\,\mathrm{d}r\right].$$

By the Law of Iterated Expectations, this can be simplified to

$$\mathrm{E}\left[\int_{T_c}^{\sigma} Af(r, Z_r)\,\mathrm{d}r \middle| \sigma > T_c\right] \mathrm{P}(\sigma > T_c).$$

But from (1.36) and the definition of T_c, we can see that for any $r > T_c$, $Af(r, x) > 0\ \forall x \in \mathbb{R}$. Therefore, stopping time $\sigma \wedge T_c$ (which differs from σ with strictly positive probability) achieves strictly lower expected loss, which is a contradiction to σ being optimal.

Finally, when $T_c \leq 0$ (which corresponds to high cost of time $\bar{c}$), we can see from (1.28) and (1.36) that it is never beneficial to wait, so any optimal stopping time is trivially bounded by 0 almost surely. □

1.A.7 Unboundedness of the decision stopping time

Lemma 1.A.4 (Fudenberg, Strack, and Strzalecki (2018))**.** *The minimal optimal decision stopping time τ^* is unbounded, i.e., $\forall t \geq 0\ \ \mathsf{P}(\tau^* > t) > 0$.*

Proof. This result basically follows from part (iii) and (iv) of Theorem 4 and footnote 22 of Fudenberg, Strack, and Strzalecki (2018). Consider first the case $X_0 = 0$. As Fudenberg, Strack, and Strzalecki (2018) show, being at the point of indifference $X_t = 0$, it is always optimal to sample for a small enough amount of time $\varepsilon > 0$ because the benefit of waiting is of order $\sqrt{\varepsilon}$ (stemming from truncated normal distribution due to transformation of the problem to a stopping of a Brownian motion) while the cost is linear. Hence, the optimal stopping boundary is non-zero for all $t \geq 0$. More specifically, Fudenberg, Strack, and Strzalecki (2018) show that the minimal optimal stopping time τ^* is $\inf\{t \geq 0 : |Z_t| \geq \alpha^2(\sigma_0^{-2} + \alpha^{-2}t)k^*(t, c, \sigma_0, \alpha)\}$, where $k^*(t, c, \sigma_0, \alpha)$ is strictly positive, strictly decreasing in t, and continuous. Therefore, for a fixed $t \geq 0$ we can find a small enough $a_t > 0$ and $\delta_t > 0$ such that the region $[0, t + \delta_t] \times (-a_t, a_t)$ is contained in the initial portion of the continuation region $\{(s, z) : s \in [0, t + \delta t], |z| \leq \alpha^2(\sigma_0^{-2} + \alpha^{-2}s)k^*(s, c, \sigma_0, \alpha)\}$. We will prove that there exists a strictly positive mass of drifts for which a Brownian

motion with either of those drifts has a strictly positive probability of staying in the region $[0, t+\delta_t] \times (-a_t, a_t)$, which will imply $\mathsf{P}(\tau^* > t) > 0$.[34]

Denote stopping time

$$\tau^{\theta}_{a_t} = \inf\{s \geq 0 : |\theta s + \alpha\sqrt{2}B_s| \geq a_t\}. \tag{1.37}$$

Consider the exponential martingale M^{η} associated with Brownian motion B defined by

$$M^{\eta}_s = \exp\left(\eta B_s - \frac{1}{2}\eta^2 s\right)$$

for $s \geq 0$ and some $\eta \in \mathbb{C}$ (Le Gall 2016, p. 50-51 and Proposition 5.11, p. 118). We can choose η of the form

$$\sqrt{\frac{\theta^2}{2\alpha^2} - u^2} - \frac{\theta}{\alpha\sqrt{2}}$$

for some $u \geq 0$ such that $\theta^2 < 2\alpha^2 u^2$. With this specific parametrization of η, we can rewrite

$$M^{\eta}_s = \exp\left(\frac{1}{2}u^2 s - \frac{\theta^2}{2\alpha^2}s - \frac{\theta}{\alpha\sqrt{2}}B_s\right)\exp\left(i\sqrt{u^2 - \frac{\theta^2}{2\alpha^2}}\left(B_s + \frac{\theta}{\alpha\sqrt{2}}s\right)\right)$$

and obtain the real part, which is also a martingale,

$$\mathrm{Re}(M^{\eta}_s) = \exp\left(\frac{1}{2}u^2 s - \frac{\theta^2}{2\alpha^2}s - \frac{\theta}{\alpha\sqrt{2}}B_s\right)\cos\left(\sqrt{u^2 - \frac{\theta^2}{2\alpha^2}}\left(B_s + \frac{\theta}{\alpha\sqrt{2}}s\right)\right).$$

If we assume, toward contradiction, that $\tau^{\theta}_{a_t} \leq t$ a.s., then by the Optional Stopping Theorem (Le Gall 2016, Corollary 3.23, p. 61) we have

$$\mathrm{E}\left[\mathrm{Re}(M^{\eta}_{\tau^{\theta}_{a_t}})\right] = \mathrm{Re}(M^{\eta}_0) = 1. \tag{1.38}$$

On the other hand, conditional on $\tau^{\theta}_{a_t} = s$, we have

$$B_{\tau^{\theta}_{a_t}} = \begin{cases} \frac{a_t}{\alpha\sqrt{2}} - \frac{\theta}{\alpha\sqrt{2}}s & \text{with probability } p(s), \\ \frac{-a_t}{\alpha\sqrt{2}} - \frac{\theta}{\alpha\sqrt{2}}s & \text{with probability } 1 - p(s), \end{cases}$$

[34]The following argument is inspired by https://math.stackexchange.com/questions/726084/does-a-brownian-motion-remain-in-any-given-open-set-for-a-given-interval-of-time.

where, for our purposes, we do not need to specify the conditional probability $p(s) \in [0,1]$. Hence, by the Law of Iterated Expectations and symmetry of the cosine function, we obtain

$$\mathrm{E}\left[\mathrm{Re}(M^{\eta}_{\tau^{\theta}_{a_t}})\right] = \cos\left(\frac{a_t}{\alpha\sqrt{2}}\sqrt{u^2 - \frac{\theta^2}{2\alpha^2}}\right) \times \mathrm{E}\left[\exp\left(\frac{1}{2}u^2\tau^{\theta}_{a_t}\right)\left(p\exp\left(-\frac{\theta}{2\alpha^2}a_t\right) + (1-p)\exp\left(\frac{\theta}{2\alpha^2}a_t\right)\right)\right].$$

Therefore, if we assume that $\tau^{\theta}_{a_t} \leq t$ a.s. and we let $u = \sqrt{\frac{\pi^2}{2}\frac{\alpha^2}{a_t^2} + \frac{\theta^2}{2\alpha^2}}$, we get $\mathrm{E}\left[\mathrm{Re}(M^{\eta}_{\tau^{\theta}_{a_t}})\right] = 0$, which is in contradiction with (1.38). Hence, $\mathsf{P}\left(\tau^{\theta}_{a_t} > t\right) > 0$. Moreover, this argument is valid for any finite $\theta \in \mathbb{R}$ because we can choose $u > 0$ accordingly.

In view of footnote 22 of Fudenberg, Strack, and Strzalecki (2018), we conclude the proof for cases $X_0 \neq 0$ by arguing that we can reach the level $-\alpha^2\sigma_0^{-2}X_0$ quickly enough. By the strong Markovian property, the argument then follows from the previous paragraph. Fortunately, an explicit expression for the probability that a Brownian motion with (arbitrarily given) drift θ, $Z_t = \theta t + \alpha\sqrt{2}B_t$, reaches a desired threshold a by a desired time T can be derived[35]

$$\exp\left(\frac{\theta a}{\alpha^2}\right)\Phi\left(-\frac{a+\theta T}{\alpha\sqrt{2T}}\right) + 1 - \Phi\left(\frac{a-\theta T}{\alpha\sqrt{2T}}\right).$$

This probability is always strictly positive. □

[35] For example, see https://galton.uchicago.edu/~yibi/teaching/stat317/2021/Lectures/Lecture25.pdf or https://math.stackexchange.com/questions/1053294/density-of-first-hitting-time-of-brownian-motion-with-drift.

1.B Deterministic stopping times

Finding the optimal deterministic stopping time

$$\inf_{t\in[0,\infty)} u(t;s,z) \tag{1.39}$$

amounts to analyzing the derivative of function u defined in (1.23).

Lemma 1.B.1.

$$\begin{aligned}\frac{\partial u(t;s,z)}{\partial t} = \bar{c} - \frac{\alpha^{-2}}{\sigma_0^{-2}+\alpha^{-2}(s+t)}\sqrt{\frac{\sigma_0^{-2}+\alpha^{-2}s}{\sigma_0^{-2}+\alpha^{-2}s+2\alpha^{-2}t}} \\ \times \frac{1}{2\pi}\exp\left(-\frac{1}{2}\frac{\sigma_0^{-2}+\alpha^{-2}(s+t)}{\sigma_0^{-2}+\alpha^{-2}s+2\alpha^{-2}t}\frac{(\sigma_0^{-2}X_0+\alpha^{-2}z)^2}{\sigma_0^{-2}+\alpha^{-2}s}\right).\end{aligned} \tag{1.40}$$

Proof. As noticed by Øksendal (2003) in (8.1.1), we can see from (1.28) that

$$\frac{\partial u(t;s,z)}{\partial t} = \mathrm{E}\left[Af(s+t, Z_t^z)\right], \tag{1.41}$$

where $Af(r,x)$ takes the form (1.36). Hence, we need to calculate

$$\mathrm{E}\left[\varphi^2\left(\frac{\sigma_0^{-2}X_0+\alpha^{-2}Z_t^z}{\sqrt{2(\sigma_0^{-2}+\alpha^{-2}(s+t))}}\right)\right].$$

By the Law of Iterated Expectations

$$\begin{aligned}&\mathrm{E}\left[\varphi^2\left(\frac{\sigma_0^{-2}X_0+\alpha^{-2}Z_t^z}{\sqrt{2(\sigma_0^{-2}+\alpha^{-2}(s+t))}}\right)\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[\varphi^2\left(\frac{\sigma_0^{-2}X_0+\alpha^{-2}Z_t^z}{\sqrt{2(\sigma_0^{-2}+\alpha^{-2}(s+t))}}\right)\middle|\theta\right]\right].\end{aligned} \tag{1.42}$$

Since

$$\left.\frac{\sigma_0^{-2}X_0+\alpha^{-2}Z_t^z}{\sqrt{2(\sigma_0^{-2}+\alpha^{-2}(s+t))}}\right|\theta \sim N(\mu,\sigma^2)$$

with

$$\mu = \frac{\sigma_0^{-2}X_0+\alpha^{-2}(z+\theta t)}{\sqrt{2(\sigma_0^{-2}+\alpha^{-2}(s+t))}}, \tag{1.43}$$

$$\sigma^2 = \frac{\alpha^{-2}t}{\sigma_0^{-2} + \alpha^{-2}(s+t)}, \tag{1.44}$$

the interior expectation in (1.42) is

$$\begin{aligned}
\mathrm{E}\left[\varphi^2\left(\frac{\sigma_0^{-2}X_0 + \alpha^{-2}Z_t^z}{\sqrt{2(\sigma_0^{-2} + \alpha^{-2}(s+t))}}\right)\middle|\theta\right] &= \int_{-\infty}^{\infty} \varphi^2(x) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \mathrm{d}x \\
&= (2\pi)^{-\frac{3}{2}}\sigma^{-1} \int_{-\infty}^{\infty} e^{-x^2 - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \mathrm{d}x.
\end{aligned} \tag{1.45}$$

We can write

$$\begin{aligned}
-x^2 - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} &= -\frac{1}{2\sigma^2}((1+2\sigma^2)x^2 - 2x\mu + \mu^2) \\
&= -\frac{1+2\sigma^2}{2\sigma^2}\left(x^2 - 2x\frac{\mu}{1+2\sigma^2} + \frac{\mu^2}{(1+2\sigma^2)^2}\right) \\
&\quad + \frac{\mu^2}{2\sigma^2(1+2\sigma^2)} - \frac{\mu^2}{2\sigma^2}.
\end{aligned} \tag{1.46}$$

Plugging (1.46) to (1.45) yields

$$\begin{aligned}
&\mathrm{E}\left[\varphi^2\left(\frac{\sigma_0^{-2}X_0 + \alpha^{-2}Z_t^z}{\sqrt{2(\sigma_0^{-2} + \alpha^{-2}(s+t))}}\right)\middle|\theta\right] \\
&= (2\pi)^{-\frac{3}{2}}\sigma^{-1} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{1+2\sigma^2}{\sigma^2}\left(x - \frac{\mu}{1+2\sigma^2}\right)^2} \mathrm{d}x \cdot e^{\frac{\mu^2}{2\sigma^2(1+2\sigma^2)} - \frac{\mu^2}{2\sigma^2}} \\
&= (2\pi)^{-\frac{3}{2}}\sigma^{-1}\sqrt{2\pi\frac{\sigma^2}{1+2\sigma^2}} \underbrace{\frac{1}{\sqrt{2\pi\frac{\sigma^2}{1+2\sigma^2}}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{1+2\sigma^2}{\sigma^2}\left(x - \frac{\mu}{1+2\sigma^2}\right)^2} \mathrm{d}x}_{=1} \cdot e^{-\frac{\mu^2}{1+2\sigma^2}} \\
&= \frac{1}{2\pi\sqrt{1+2\sigma^2}} e^{-\frac{\mu^2}{1+2\sigma^2}}.
\end{aligned} \tag{1.47}$$

By Lemma 1.A.1, the beliefs about θ at (s, z) are

$$N(X_s, \sigma_s^2)$$

with

$$\begin{aligned}
X_s &= \frac{\sigma_0^{-2}X_0 + \alpha^{-2}z}{\sigma_0^{-2} + \alpha^{-2}s}, \\
\sigma_s^2 &= \frac{2}{\sigma_0^{-2} + \alpha^{-2}s}.
\end{aligned}$$

Hence, we can insert (1.47) back to (1.42) and calculate the outer expectation explicitly. Let us focus on the part in (1.47) that depends on θ

$$\mathrm{E}\left[e^{-\frac{\mu^2}{1+2\sigma^2}}\right] = \frac{1}{\sqrt{2\pi\sigma_s^2}} \int_{-\infty}^{\infty} e^{-\left(\frac{\mu^2}{1+2\sigma^2}+\frac{(\theta-X_s)^2}{2\sigma_s^2}\right)} \mathrm{d}\theta. \tag{1.48}$$

We can write

$$\frac{\mu^2}{1+2\sigma^2} + \frac{(\theta - X_s)^2}{2\sigma_s^2} = (2\sigma_s^2(1+2\sigma^2))^{-1}[(\theta^2 - 2\theta X_s + X_s^2)(1+2\sigma^2) + 2\sigma_s^2\mu^2]. \tag{1.49}$$

By expanding μ according to (1.43) and denoting

$$\sigma_{s+t}^2 = \frac{2}{\sigma_0^{-2} + \alpha^{-2}(s+t)},$$

we can continue rewriting (1.49) as

$$\begin{aligned}(2\sigma_s^2(1+2\sigma^2))^{-1}\big[&\theta^2(1+2\sigma^2+\frac{1}{2}\sigma_s^2\sigma_{s+t}^2\alpha^{-4}t^2) \\ &- 2\theta(X_s(1+2\sigma^2) - \frac{1}{2}\sigma_s^2\sigma_{s+t}^2(\sigma_0^{-2}X_0 + \alpha^{-2}z)\alpha^{-2}t) \\ &+ X_s^2(1+2\sigma^2) + \frac{1}{2}\sigma_s^2\sigma_{s+t}^2(\sigma_0^{-2}X_0 + \alpha^{-2}z)^2\big].\end{aligned} \tag{1.50}$$

Using $\sigma^2 = \sigma_{s+t}^2\alpha^{-2}t/2$ and $X_s = \sigma_s^2(\sigma_0^{-2}X_0 + \alpha^{-2}z)/2$, we can simplify the terms in (1.50) and introduce term K to simplify further calculations

$$\begin{aligned}K &:= 1 + 2\sigma^2 + \frac{1}{2}\sigma_s^2\sigma_{s+t}^2\alpha^{-4}t^2 = 1 + \sigma_{s+t}^2\alpha^{-2}t(1 + \frac{1}{2}\sigma_s^2\alpha^{-2}t) \\ &= 1 + \frac{2\alpha^{-2}t}{\sigma_0^{-2} + \alpha^{-2}(s+t)}\left(1 + \frac{\alpha^{-2}t}{\sigma_0^{-2} + \alpha^{-2}s}\right) = \frac{\sigma_0^{-2} + \alpha^{-2}s + 2\alpha^{-2}t}{\sigma_0^{-2} + \alpha^{-2}s},\end{aligned} \tag{1.51}$$

$$X_s(1+2\sigma^2) - \frac{1}{2}\sigma_s^2\sigma_{s+t}^2(\sigma_0^{-2}X_0 + \alpha^{-2}z)\alpha^{-2}t = X_s, \tag{1.52}$$

$$\begin{aligned}&X_s^2(1+2\sigma^2) + \frac{1}{2}\sigma_s^2\sigma_{s+t}^2(\sigma_0^{-2}X_0 + \alpha^{-2}z)^2 \\ &\quad = X_s^2(1 + \sigma_{s+t}^2\alpha^{-2}t + 2\sigma_s^{-2}\sigma_{s+t}^2) = X_s^2[1 + \sigma_{s+t}^2(\underbrace{\alpha^{-2}t + 2\sigma_s^{-2}}_{2\sigma_{s+t}^{-2}})] = 3X_s^2.\end{aligned} \tag{1.53}$$

Then, we can continue with (1.50)

$$(2\sigma_s^2(1+2\sigma^2))^{-1}\left[K\theta^2 - 2\theta X_s + 3X_s^2\right]$$

$$= (2\sigma_s^2(1+2\sigma^2))^{-1} K \left[\theta^2 - 2\theta\frac{X_s}{K} + \frac{X_s^2}{K^2} - \frac{X_s^2}{K^2} + \frac{3X_s^2}{K}\right]$$
$$= \frac{K}{2\sigma_s^2(1+2\sigma^2)}\left[\left(\theta - \frac{X_s}{K}\right)^2 - \frac{X_s^2}{K^2} + \frac{3X_s^2}{K}\right].$$

Plugging this back to (1.48) yields

$$\begin{aligned}
\mathrm{E}\left[e^{-\frac{\mu^2}{1+2\sigma^2}}\right] &= \exp\left(-\frac{K}{2\sigma_s^2(1+2\sigma^2)}\left(-\frac{X_s^2}{K^2} + \frac{3X_s^2}{K}\right)\right) \\
&\quad \times \frac{1}{\sqrt{2\pi\sigma_s^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\frac{\sigma_s^2(1+2\sigma^2)}{K}}\left(\theta - \frac{X_s}{K}\right)^2\right) \mathrm{d}\theta \\
&= \exp\left(\frac{1}{2\sigma_s^2(1+2\sigma^2)}\left(\frac{X_s^2}{K} - 3X_s^2\right)\right)\sqrt{\frac{1+2\sigma^2}{K}} \\
&\quad \times \underbrace{\frac{1}{\sqrt{2\pi\sigma_s^2\frac{1+2\sigma^2}{K}}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\frac{\sigma_s^2(1+2\sigma^2)}{K}}\left(\theta - \frac{X_s}{K}\right)^2\right) \mathrm{d}\theta}_{=1} \\
&= \exp\left(\frac{1}{2\sigma_s^2(1+2\sigma^2)}\left(\frac{X_s^2}{K} - 3X_s^2\right)\right)\sqrt{\frac{1+2\sigma^2}{K}}. \qquad (1.54)
\end{aligned}$$

The following steps lead to a simplification of expression (1.54):

$$\begin{aligned}
2\sigma_s^2(1+2\sigma^2) &= 4\frac{\sigma_0^{-2} + \alpha^{-2}s + 3\alpha^{-2}t}{(\sigma_0^{-2} + \alpha^{-2}s)(\sigma_0^{-2} + \alpha^{-2}(s+t))}, \\
1 - 3K &= \frac{-2(\sigma_0^{-2} + \alpha^{-2}s + 3\alpha^{-2}t)}{\sigma_0^{-2} + \alpha^{-2}s}, \\
\frac{1-3K}{K} &= \frac{-2(\sigma_0^{-2} + \alpha^{-2}s + 3\alpha^{-2}t)}{\sigma_0^{-2} + \alpha^{-2}s + 2\alpha^{-2}t}, \\
\frac{1}{2\sigma_s^2(1+2\sigma^2)}\frac{1-3K}{K}X_s^2 &= -\frac{1}{2}\frac{\sigma_0^{-2} + \alpha^{-2}s + \alpha^{-2}t}{\sigma_0^{-2} + \alpha^{-2}s + 2\alpha^{-2}t}\frac{(\sigma_0^{-2}X_0 + \alpha^{-2}z)^2}{\sigma_0^{-2} + \alpha^{-2}s}. \qquad (1.55)
\end{aligned}$$

Therefore, by putting together (1.36), (1.41), (1.42), (1.47), (1.51), (1.54), and (1.55), we obtain

$$\begin{aligned}
\frac{\partial u(t;s,z)}{\partial t} &= \bar{c} - \frac{\alpha^{-2}}{\sigma_0^{-2} + \alpha^{-2}(s+t)}\sqrt{\frac{\sigma_0^{-2} + \alpha^{-2}s}{\sigma_0^{-2} + \alpha^{-2}s + 2\alpha^{-2}t}} \\
&\quad \times \frac{1}{2\pi}\exp\left(-\frac{1}{2}\frac{\sigma_0^{-2} + \alpha^{-2}(s+t)}{\sigma_0^{-2} + \alpha^{-2}s + 2\alpha^{-2}t}\frac{(\sigma_0^{-2}X_0 + \alpha^{-2}z)^2}{\sigma_0^{-2} + \alpha^{-2}s}\right).
\end{aligned}$$

□

1.C Confidence in the Wald model

Fudenberg, Strack, and Strzalecki (2018) also study a variant of the model with a priori known difference between the options $\theta_d := |\theta_l - \theta_r|$ with the only uncertainty about which option is the best. This corresponds to the binomial prior about θ:

$$\theta = \begin{cases} \theta_d & \text{with probability } \mu_0 \in (0,1), \\ -\theta_d & \text{with probability } 1-\mu_0. \end{cases} \tag{1.56}$$

This model is the continuous-time version of the model of Wald (1947). As Fudenberg, Strack, and Strzalecki (2018) write, Shiryaev (2007) provides a solution to this problem.

Theorem 1.C.1 (Shiryaev (2007))**.** *For the model with the binomial prior, there exists $k > 0$ such that the minimal optimal stopping time is $\hat{\tau} = \inf\{t \geq 0 : |l_t| = k\}$, where $l_t = \log\left(\frac{p_t}{1-p_t}\right)$ and $p_t = \mathsf{P}(\theta = \theta_d | \mathcal{F}_t)$.*

Since p_t can be seen as belief confidence (in the left option), this solution is often interpreted as a decision maker committing to a desired level of confidence ex ante. Since there is no further deliberation in this model, the predicted (one-stage) decision confidence is constant $\frac{e^k}{1+e^k}$.

Given the interest of this paper, we can ask whether there is a scope for two-stage confidence in this model. Motivated by calculations (1.10), we will analyze the unconstrained stopping problem

$$\inf_{\tau' \in \mathcal{T}} \mathrm{E}\left[p_{\tau'}(1-p_{\tau'}) + \bar{c}\tau'\right]. \tag{1.57}$$

The belief process $\{p_t\}_{t\geq 0}$ solves the stochastic differential equation

$$\mathrm{d}p_t = p_t(1-p_t)\frac{\sqrt{2}}{\alpha}\mathrm{d}W_t,$$

where $\{W_t\}_{t\geq 0}$ is a Brownian motion with respect to $\mathcal{F}_t$ (Morris and Strack 2019, Lemma 1). By applying the differential operator of the process $\{(s+t, p_t^p)\}_{t\geq 0}$ to the function $(t,p) \mapsto p(1-p) + \bar{c}t$ and inspecting where it is negative, we obtain a candidate stopping

region in the (t,p)-space

$$\mathring{C}_W = \left\{(t,p) \in [0,\infty) \times [0,1] : p(1-p) > \alpha\sqrt{\frac{\bar{c}}{2}}\right\}.$$

If $\bar{c} > \frac{1}{8\alpha^2}$, this region is empty and it is always optimal to stop immediately. Otherwise this region implicitly prescribes constant boundaries for p_t, $0 \leq p_* \leq \frac{1}{2} \leq p^* \leq 1$.

We will show that the optimal continuation region takes the same form as $\mathring{C}_W$, i.e.,

$$C_W := \left\{(t,p) \in [0,\infty) \times [0,1] : \underline{p} < p < \bar{p}\right\} \tag{1.58}$$

with the boundaries $\underline{p} \leq p_* \leq \frac{1}{2} \leq p^* \leq \bar{p}$ (since $\mathring{C}_W \subseteq C_W$). To do that, we follow a similar approach as Øksendal (2003, p. 210).

Denote the value function

$$V(s,p) = \inf_{\tau \in \mathcal{T}} \mathrm{E}\left[p_\tau^p(1-p_\tau^p) + \bar{c}(\tau+s)\right],$$

where

$$p_t^p = p + \int_0^t p_r(1-p_r)\frac{\sqrt{2}}{\alpha}\mathrm{d}W_r.$$

Since

$$\begin{aligned} V(s-t_0,p) &= \inf_{\tau \in \mathcal{T}} \mathrm{E}\left[p_\tau^p(1-p_\tau^p) + \bar{c}(\tau+s-t_0)\right] \\ &= \inf_{\tau \in \mathcal{T}} \mathrm{E}\left[p_\tau^p(1-p_\tau^p) + \bar{c}(\tau+s)\right] - \bar{c}t_0 \\ &= V(s,p) - \bar{c}t_0, \end{aligned}$$

we can show that C_W is invariant in time:

$$\begin{aligned} C_W + (t_0,0) &= \{(t_0+t,p) : (t,p) \in C_W\} \\ &= \{(s,p) : (s-t_0,p) \in C_W\} \\ &= \{(s,p) : V(s-t_0,p) < p(1-p) + \bar{c}(s-t_0)\} \\ &= \{(s,p) : V(s,p) - \bar{c}t_0 < p(1-p) + \bar{c}(s-t_0)\} \\ &= \{(s,p) : V(s,p) < p(1-p) + \bar{c}s\} = C_W. \end{aligned}$$

Hence, C_W must take the form (1.58).

This analysis reveals that if the parameters $\bar{c}, c, \alpha, \mu_0$ are such that $\frac{e^k}{1+e^k} \geq \bar{p}$, confidence will always be one-stage with the resulting confidence level $\frac{e^k}{1+e^k}$; otherwise confidence will always be two-stage with the two potential levels $\underline{p}$ (for error recognition) and $\bar{p}$ (for correctness confirmation).

1.D Numerical solution

In this section, we describe the numerical methods for finding the optimal decision and confidence stopping boundaries ∂C_D and ∂C_C, respectively.

The decision boundary is found by building on the code and descriptions of Fudenberg, Strack, and Strzalecki (2018) from their paper, Online Appendix, and replication folder accessible at OPENICPSR. In particular, script `h_patching.m`[36] imports all csv files (except for `h-30-100.csv` because they write in Section 5 of `readme.pdf` that it is just a mistake) from the subfolder 'solver/mat' of their replication folder and patches them together as described in Section 4.1.2 of their Online Appendix. Moreover, script `h_patching.m` also (i) reduces the size of the resulting vector by dropping unnecessary points[37] and (ii) rectifies monotonicity as mentioned in Footnote 2 of their Online Appendix.[38] The resulting vector representing their h function is saved and used by function `opt_dec_bound.m` to calculate the optimal decision boundary, which Fudenberg, Strack, and Strzalecki (2018) denote by b^* and define in Theorem 4 (page 3661), with the generalization to non-symmetric prior means in their Footnote 22 (page 3662).[39] Function `opt_dec_bound.m` thus takes as input characteristics/parameters of an individual (c, σ_0, α, X_0) and the h vector, and outputs upper and lower portions of the decision stopping boundary.

The confidence boundary is computed by function `conf_bound.m`, which takes as input characteristics/parameters of an individual ($\bar{c}$, σ_0, α, X_0) and outputs a list of points in the (t, z)-space. The computation is performed by backward induction. Denote by $\mathcal{U}$ the numerical approximation of the value function U defined in (1.24).

$\mathcal{U}$ is computed as follows. Fix a space grid by choosing sufficiently large (in absolute value) $\bar{z} > 0$ and $\underline{z} < 0$ and a sufficiently small $\Delta_z > 0$: $\zeta = \{\underline{z}, \underline{z} + \Delta_z, \ldots, \bar{z} - \Delta_z, \bar{z}\}$.[40]

[36] The codes are available upon request.

[37] The original resulting vector has more than 95% of useless constant parts, which we drop and use linear interpolation instead to speed up further computations.

[38] 9 points are redefined.

[39] See also Lemma O.4 of their Online Appendix for details of these calculations.

[40] Δ_z and Δ_t are set to 0.1 (unless the time bound from Proposition 1.A.3 is smaller than 5; then a smaller Δ_t is chosen). Initially, $\bar{z}$ and $\underline{z}$ are set ad hoc to $50 - \alpha^2 \sigma_0^{-2} X_0$ and $-50 - \alpha^2 \sigma_0^{-2} X_0$, respectively. During the backward induction computation, if these initial $\bar{z}$ and $\underline{z}$ become insufficient (the computed

Start at T slightly above the bound from Proposition 1.A.3 where it is surely optimal to stop for any $z \in \mathbb{R}$, i.e., set $\mathcal{U}(T, z) = f(T, z)$. Going backwards in time by Δ_t, suppose we already computed $\mathcal{U}(s, z),\ z \in \zeta$, for all $s \in \{t + \Delta_t, \ldots, T\}$. At time point t for space point $z \in \zeta$, we want to compare $f(t, z)$ and

$$\mathrm{E}_{(t,z)}\left[U(t + \Delta_t, Z^z_{\Delta_t})\right] = \int_{-\infty}^{\infty} U(t + \Delta_t, x) g_{(t,z)}(x)\,\mathrm{d}x, \tag{1.59}$$

where $g_{(t,z)}$ is the pdf of the normal distribution

$$N\left(z + \frac{\sigma_0^{-2} X_0 + \alpha^{-2} z}{\sigma_0^{-2} + \alpha^{-2} t}\Delta_t, 2\alpha^2 \Delta_t + \frac{2}{\sigma_0^{-2} + \alpha^{-2} t}\Delta_t^2\right).$$

Denote the cdf of this distribution $G_{(t,z)}$. To approximate the expected value (1.59), use a constant approximation of $U(t + \Delta_t, x)$ below $\underline{z}$ and above $\bar{z}$[41] and the trapezoidal rule on ζ

$$\begin{aligned}
&\int_{-\infty}^{\infty} U(t + \Delta_t, x) g_{(t,z)}(x)\,\mathrm{d}x \approx U(t + \Delta_t, \underline{z}) G_{(t,z)}(\underline{z}) \\
&\quad + \sum_{x \in \{\underline{z}, \underline{z} + \Delta_z, \ldots, \bar{z} - \Delta_z\}} \frac{1}{2}[U(t + \Delta_t, x) g_{(t,z)}(x) + U(t + \Delta_t, x + \Delta_z) g_{(t,z)}(x + \Delta_z)]\Delta_z \\
&\quad + U(t + \Delta_t, \bar{z})(1 - G_{(t,z)}(\bar{z})) \\
&\quad =: \mathcal{E}_{(t,z)}.
\end{aligned}$$

Finally, set

$$\mathcal{U}(t, z) = \min\{f(t, z), \mathcal{E}_{(t,z)}\}.$$

The numerically computed continuation region is

$$\{(t, z) : \mathcal{U}(t, z) < f(t, z)\}.$$

More precisely, for a given $t \in \{0, \Delta_t, \ldots, T\}$, we find the extreme points $z \in \zeta$ for which $\mathcal{U}(t, z) < f(t, z)$ and consider them to be on the boundary of the confidence continuation region.

---

continuation region approaches $0.8\bar{z}$ from below or $0.8\underline{z}$ from above), the backward induction repeatedly restarts anew with a larger range of bounds until it is sufficient.

[41] Below $\underline{z}$ and above $\bar{z}$, the MSE part of the loss function is almost zero, so $U(t + \Delta_t, x)$ flattens out to approximately $\bar{c}(t + \Delta_t)$ beyond these points.

1.E Empirical evaluation of models

Table 1.E.1: Evaluation of models using empirical patterns of Moran et al. (2015)

	Empirical pattern	Explanation	FSS	2SS
1.	Speed-accuracy trade-off	Higher error rate under time pressure	✓	✓
2.	Slow/fast errors	Error choices can be slower or faster than correct choices	Slow errors	Slow errors
3.	Negative correlation of confidence and difficulty	Positive correlation of confidence and stimulus discriminability	✓	✓
4.	Negative correlation of decision time and confidence		✓	✓
5.	Lower confidence under time pressure		✓	✓
6.	Positive confidence resolution	Higher confidence in correct decisions	✓	✓
7.	Increased confidence resolution under time pressure	Difference in confidence between correct and error choices is higher under time pressure	×	✓
8.1.	Positive correlation of RT2 and difficulty	Negative correlation of RT2 and stimulus discriminability	–	✓
8.2.	Lower RT2 in correct choices		–	✓
8.3.	Negative correlation of RT2 and confidence		–	✓
8.4.	Positive correlation of RT2 and RT		–	×
9.	Decreased confidence resolution for difficult decisions	Difference in confidence between correct and error choices is lower under lower stimulus discriminability	✓	✓
10.	Higher RT2 for correct choices and lower RT2 for errors under higher difficulty		–	✓

FSS is the model of Fudenberg, Strack, and Strzalecki (2018) with confidence determined at decision stopping. 2SS is the model proposed in this paper. We tried the following parameters: (i) $c = 0.06, \sigma_0 = 1, \alpha = 2, X_0 = 0$ with $\bar{c} = 0.02$ for 2SS model ("typical" subject as in Figure 1.G.1), (ii) $c = 0.02, \sigma_0 = 1.8, \alpha = 2, X_0 = 0$ with $\bar{c} = 0.012$ for 2SS model ("atypical" subject as in Figure 1.2). The results are based on 100,000 simulated trials for each model and set of parameters. We tried θ distributed according to (a) the agent's prior $N(X_0, 2\sigma_0^2)$, (b) $N(0, 1)$. Discriminability is measured by $|\theta|$. RT is response time, RT2 is interjudgement time. Time pressure effect was analyzed by decreasing c to 0.05 in (i) and 0.01 in (ii).

1.F Literature: economic motivation for decision confidence

- Enke and Graeber (2022) demonstrate the economic relevance of decision confidence in predicting behavioral biases.
- Enke, Graeber, and Oprea (2022) show that confidence modulates attenuation of behavioral biases in aggregate through self-selection into institutions.
- Folke et al. (2016) show that "an explicit representation of confidence is harnessed for subsequent changes of mind." This may be relevant for markets with the possibility of changing one's mind, e.g., return policies in online shopping or cancellation insurance in airlines or races.
- Van den Berg et al. (2016) show that people exploit confidence in previous choices to adjust their termination mechanism in subsequent decisions. This may be relevant for projects, which are typically composed of several subsequent decisions directed towards an overarching goal. Moreover, it is well known that demand for many products is derived from demand for other products; however, the demand for the "later" products may also be affected by the mere confidence in the "earlier" products, e.g., demand for electricity may be affected by one's decision confidence in a purchase of an appliance if one cares about its energy efficiency.
- Purcell and Kiani (2016) show that people use confidence to modulate the adjustment of their strategy when facing negative feedback, i.e., confidence disambiguates between two sources of errors: bad strategy vs. bad information. This suggests an important role of confidence in reinforcement learning and the Credit Assignment Problem in particular.
- Yin, Mitra, and Zhang (2016) document that consumers exhibit confirmation bias in evaluating online reviews and this bias is amplified by higher confidence in their initial beliefs. Hence, confidence plays a role in information acquisition (Schulz, Fleming, and Dayan 2021).
- Confidence is often communicated to others and it has a strong influence on their decisions (Vullioud et al. 2017; Shea et al. 2014; Sah, Moore, and MacCoun 2013; Brewer and Burke 2002).

- Artiga González, Capozza, and Granic (2022) show that voters change their policy preferences (in line with the supported candidate) after merely expressing their support. We suspect that confidence may play a role in this process of formation of policy preferences and polarization.

1.G Additional figures

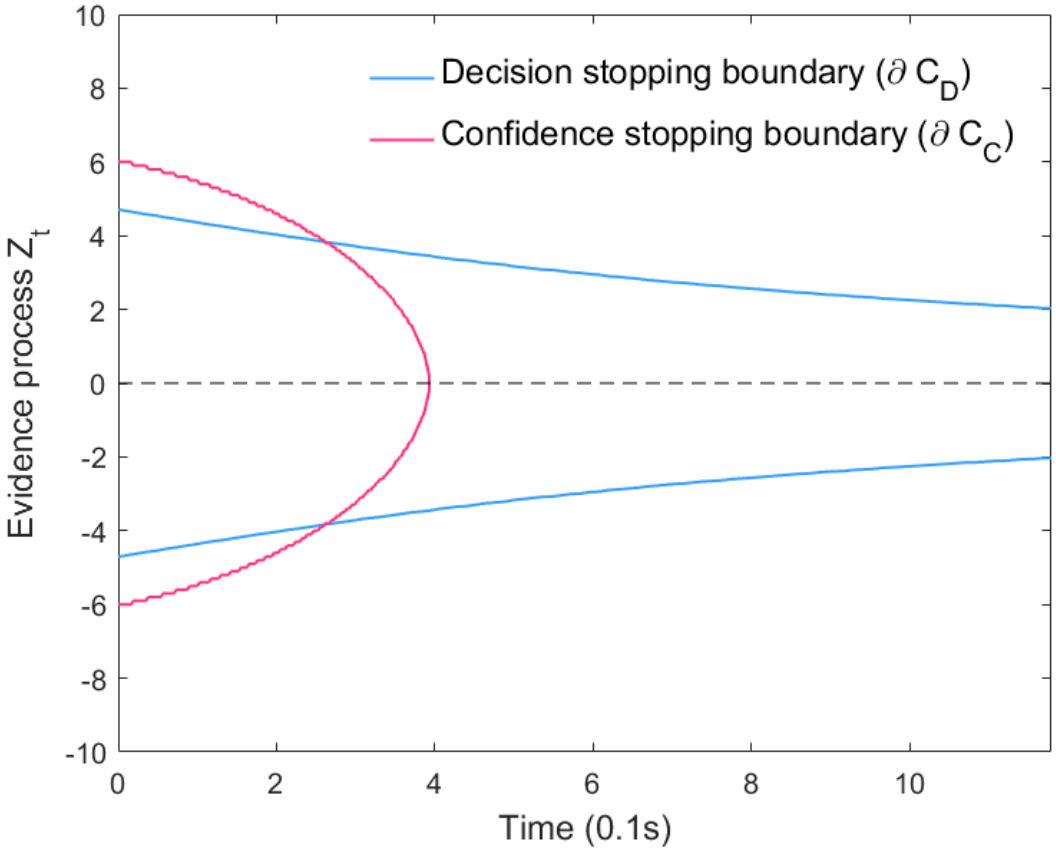


Figure 1.G.1: An example of decision and confidence stopping boundaries computed numerically (see Appendix 1.D) for parameters $c = 0.06, \alpha = 2, \sigma_0 = 1, X_0 = 0$ (corresponding to Subject 47 in Table 4 in the Online Appendix of Fudenberg et al. [2018]) and $\bar{c} = 0.02$.

Chapter 2

Form of Preference Misalignment Linked to State-Pooling Structure in Bayesian Persuasion

Abstract

Rastislav Rehák and Maxim Senkov[1]

We study a Bayesian persuasion model in which the state space is finite, the sender and the receiver have state-dependent quadratic loss functions, and their disagreement regarding the preferred action is of arbitrary form. This framework enables us to focus on the understudied sender's trade-off between the informativeness of the signal and the concealment of the state-dependent disagreement about the preferred action. In particular, we study which states are pooled together in the supports of posteriors of

[1]This chapter is joint work based on Rehák, R., and Senkov, M. (2021) "Form of Preference Misalignment Linked to State-Pooling Structure in Bayesian Persuasion," CERGE-EI Working Paper Series No. 708. This project was supported by Charles University GAUK project No. 666420 and by the H2020-MSCA-RISE project GEMCLIME-2020 GA No. 681228. This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreements No. 101002898 and No. 770652). We acknowledge financial support by the Lumina Quaeruntur fellowship (LQ300852101, Challenges to Democracy) of the Czech Academy of Sciences. We thank Jan Zápal, Ole Jann, Inés Moreno de Barreda, Fedor Sandomirskiy, Ludmila Matysková, Filip Matějka, and the conference audience at GAMES 2020/1 for useful comments.

the optimal signal. We provide an illustrative graph procedure that takes the form of preference misalignment and outputs potential representations of the state-pooling structure. Our model provides insights into situations in which the sender and the receiver care about two different but connected issues, for example, the interaction of a political advisor who cares about the state of the economy with a politician who cares about the political situation.

Keywords: Bayesian Persuasion, Strategic State Pooling, Preference Misalignment, Graph Procedure
JEL Codes: D82, D83

2.1 Introduction

Bayesian persuasion, pioneered by Kamenica and Gentzkow (2011), studies strategic disclosure of information when the sender controls the information environment (called *signal*) and the receiver controls the choice of action to be taken. As a review by Kamenica (2019) suggests, this literature has provided many extensions of the original model of Kamenica and Gentzkow (2011) with interesting qualitative insights. However, full characterization of the optimal signal is generally difficult even in the original model. There has been little progress on this front, and it has been limited to a small number of special cases.[2]

We contribute to this literature by studying a special case of the original model that has received little attention—a Bayesian persuasion model in which both the sender and the receiver have *state-dependent preferred actions*. We characterize a qualitative property of the optimal signal called *state-pooling structure*, which describes pools of states that cannot be discerned from one another by the optimal signal. Specifically, we ask how the structure of state-dependent preference misalignment affects the state-pooling structure of the optimal signal.

To illustrate the main point of this paper, we present an example of a politician (receiver, he) and his advisor (sender, she). They both wish to implement some level of government spending $a \in \mathbb{R}$ that is adapted to the current economic situation captured by GDP per

[2]We return to this point in the discussion of related literature in Section 2.2.

capita y, which takes one of three possible values: 1, 2, or 3. However, they each have a different vision of optimal spending as a function of GDP per capita. The advisor's payoff is $u_S(a, y) = -(a - \omega(y))^2$ and the politician's payoff is $u_R(a, y) = -(a - \rho(y))^2$, where $\omega(y)$ and $\rho(y)$ represent the preferred spending of the advisor and the politician in state y, respectively. The advisor designs an investigation (a signal) that can inform the politician about the realization of GDP per capita. She does that strategically to influence the spending choice of the politician. We are interested in how the structure of this signal depends on the form of misalignment between the advisor's and politician's preferences captured by ω and $\rho(\omega)$, respectively (we assume that $\omega(y)$ is a bijection, so that we can capture the misalignment directly by $\rho(\omega)$).

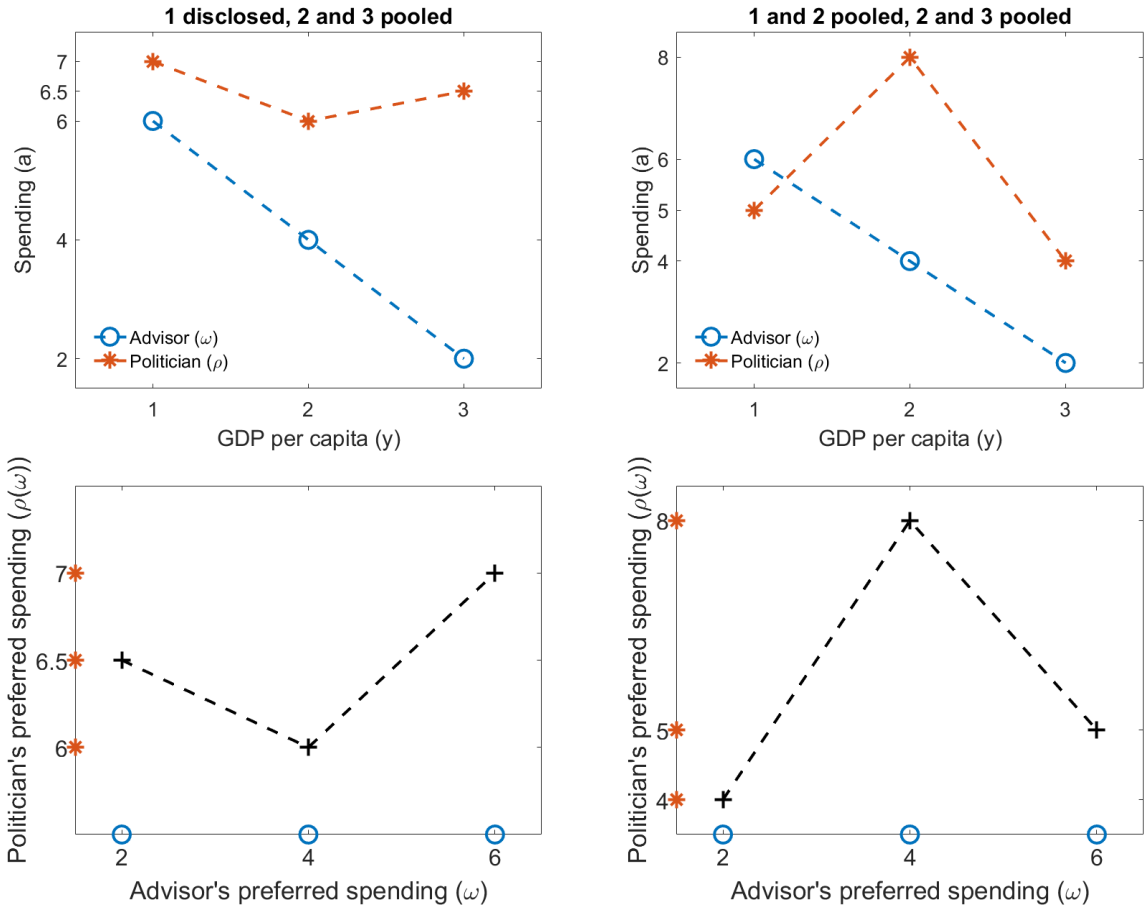


Figure 2.1: The form of disagreement between the advisor (ω) and the politician (ρ) matters for the structure of the optimal signal:
left plots: the advisor fully reveals state 1 and pools states 2 and 3 together;
right plots: the advisor pools states 1 and 2 together and states 2 and 3 together
(note: we consider only three levels of GDP per capita; the lines are drawn only for clarity of the pictures)

Figure 2.1 illustrates how the form of disagreement between the advisor's and politician's

preferred spending influences the structure of the optimal signal.[3] In the case presented *in the left plot*, the advisor's optimal signal fully reveals whether the state of the economy is low or not, i.e., one of the two outcomes of her investigation fully reveals the low state and the other leaves the politician uncertain about the high and middle states—we say they are pooled together. Intuitively, both the advisor and the politician want the highest spending in the low state, so their goals are aligned in this state and the advisor wants to reveal it perfectly. However, they disagree about whether the spending should be higher in the middle or high state, so the advisor wants to attenuate this disagreement by pooling these two states together. In the case presented *in the right plot*, the advisor's optimal signal reveals whether the economy is above or below average, i.e., one of the two outcomes of her investigation pools the low and middle states, while the other pools the middle and high states. Intuitively, the advisor and the politician disagree about whether the spending should be higher in the low or middle state, so the advisor wants to attenuate this disagreement by pooling these two states together. However, they both agree that the spending should be higher in the middle state than in the high state, but the politician prefers a greater spending difference between these two states than the advisor. Therefore, the advisor wants to moderate the politician's actions by pooling these two states together.

In Section 2.3, we describe our model. We use the Bayesian persuasion framework of Kamenica and Gentzkow (2011) with one-dimensional finite state space—the sender's preferred action. Both the sender and the receiver have quadratic loss functions with bliss points depending on the state of the world. The structure of misalignment is captured by function ρ mapping the state of the world (the sender's preferred action) to the receiver's preferred action. The case of linear ρ with slope 1 corresponds to the benchmark of perfect alignment.[4] We do not impose any requirements on this function and we analyze the role of its shape for the qualitative structure of the optimal signal in terms of state pooling.

In Section 2.4, we present general results on the pooling structure of the optimal signal. The patterns of pooling are driven by the sender's trade-off between (i) the informative-

[3]The structures of the optimal signals for the two cases considered in Figure 2.1 are derived using results from Section 2.6.

[4]A state-independent intercept does not affect the choice of the signal because it is a "sunk cost" for the sender.

ness of the signal, which leads to better adaptation of the action to the state of the world in states of alignment, and (ii) the revelation of the realized mismatch of the sender's and receiver's preferred actions, which drives the action of the receiver away from the sender's preferred action. First, we show that the sender generically benefits from revealing some information. The only cases in which non-disclosure is optimal are when ρ is linear with a slope sufficiently different from 1. Second, we show that linear ρ with a moderate slope (namely, between zero and two) leads to full disclosure. Third, we demonstrate that the optimal signal does not induce an interior belief (except in cases of non-disclosure).

In Section 2.5, we propose a simple graph procedure to characterize the optimal structure of state pooling for a given ρ. This procedure consists of an analysis of ρ on pairs of states and a test of pooling of more than two states. The crucial element of this procedure is the slope of ρ between pairs of states, which plays the role of an index of misalignment—if it is too high (disagreement about magnitude) or lower than zero (disagreement about order), then it indicates space for pooling; otherwise, it indicates space for separation.

In Section 2.6, we provide a full characterization of the state-pooling structure in the case of three states of the world. The state-pooling structure is completely pinned down by the shape of ρ except for the case in which ρ has a slope sufficiently different from 1 for each of the three pairs of states. In that case, the choice of a particular state-pooling structure depends both on the shape of ρ and the prior.

2.2 Related literature

First, we relate our work to the *Bayesian persuasion* literature. The most relevant results from the seminal paper by Kamenica and Gentzkow (2011) are (i) conditions for full disclosure or non-disclosure in the general form and (ii) comparative statics of more aligned preferences. Regarding point (i), we go beyond these two "corner" cases for the optimal signal, similarly as in the recent studies of Arieli et al. (2020) and Kolotilin and Wolitzky (2020). We discuss the connection of our work to Kolotilin and Wolitzky (2020) in more detail later in this section. Regarding point (ii), we perform a different exercise with preference misalignment: we fix the preferences and analyze how the structure of preference misalignment is related to the structure of state pooling of the optimal signal.

The methodological progress in Bayesian persuasion on the front of providing a general

characterization of the structure of the optimal signal has been scarce. First, with two or three states of the world, concavification provides an insightful graphical method of solving the sender's problem (Kamenica and Gentzkow 2011). Second, when the sender's utility depends only on the expected state, the "Rothschild-Stiglitz approach" (Gentzkow and Kamenica 2016) and linear programming methods (Kolotilin 2018; Dworczak and Martini 2019) have been used to solve these problems. However, we are interested in situations with the *sender's state-dependent preferred* action and the role of the structure of preference misalignment, where these methods do not deliver immediate answers. We propose a new concavification-based approach of characterizing the state-pooling structure of the optimal signal.

The closest paper to ours is Kolotilin and Wolitzky (2020). However, we differ along several directions, and our paper can be viewed as complementary to theirs. First, their sender prefers higher actions independently of the state, but experiences state-dependent loss from mismatching the preferred action. In contrast, our sender has state-dependent preferred actions, but her loss from mismatching the preferred action is state-independent. Second, their receiver prefers higher actions in higher states; we do not impose this assumption. Third, they provide sufficient (and "almost necessary") conditions for special patterns of "assortative" disclosure. However, they do not provide a procedure for finding the pooling structure of the optimal signal explicitly, and they avoid characterization of more complicated patterns. In contrast, we work in a more specialized quadratic setting and do not restrict ourselves to characterization of specific (pairwise) pooling structures. Instead, we propose a general procedure for finding the pooling structure. Finally, the mechanisms driving the results in the two papers are different: in Kolotilin and Wolitzky (2020), the information does not have value for the sender alone, so state pooling emerges from pure persuasion concerns, while state pooling in our model is driven by the interplay of the sender's incentives to disclose the state and to hide misalignment.

Two other related papers in Bayesian persuasion literature are Alonso and Camara (2016) and Galperti (2019). Similar to our paper, both rely on the concavification technique to obtain insights regarding the optimal signal. Alonso and Camara (2016) consider the standard Bayesian persuasion model, but assume that the sender and the receiver have heterogeneous prior beliefs. While the sender in Alonso and Camara (2016) uses the variation of the difference between the sender's and receiver's prior beliefs across the states of the world to design the optimal disclosure, our sender uses the variation

in the misalignment of the sender's and receiver's bliss points across the states of the world.[5] Galperti (2019) considers the standard Bayesian persuasion model in which the sender and the receiver have a special type of heterogeneous prior beliefs: the receiver attaches zero probability to some states that are perceived with positive probability by the sender. While we restrict attention to a sender with state-dependent bliss actions and study the general patterns of state pooling, Galperti (2019) makes weaker assumptions about preferences and focuses on patterns of pooling of the states that have a priori zero probability for the receiver.

Second, the results of our study are connected to the *literature on persuasion games*, in which the sender chooses how to disclose her private verifiable information regarding the state of the world. Milgrom (1981) and Milgrom and Roberts (1986) analyze the conventional model of a persuasion game and establish the result on "unraveling" of the sender's private information leading to full disclosure. Dye (1985) and Shin (1994) study state pooling in a similar game but with (second-order) uncertainty of the receiver about whether the sender actually has some private information or not. Seidmann and Winter (1997) analyze a persuasion game in which the sender has state-dependent preferred actions, and they demonstrate that the "unraveling" result still holds. The combination of these two features—second-order uncertainty and state-dependent preferred actions—has been studied in a small number of recent papers. The closest paper to ours is Hummel, Morgan, and Stocken (2018), in which unraveling does not occur due to the presence of the receiver's second-order uncertainty. In the Bayesian persuasion model that we study, the sender's disclosure mechanism serves a similar role to the one in Hummel, Morgan, and Stocken (2018): the sender moderates the receiver's actions via pooling of the states for which the sender's bliss-point line is sufficiently flat relative to that of the receiver.

Finally, Miura (2018) studies how pooling equilibria can be characterized based on a procedure that uses a *masquerade graph* introduced in Hagenbach, Koessler, and Perez-Richet (2014). In his procedure, a pool of states is formed by the types of the sender who are mutually interested in masquerading, i.e., being perceived by the receiver as some other type in the pool. In spirit, this resembles the procedure for discovery of the state-pooling structure we introduce: a masquerade edge between two nodes (types) in

[5]They demonstrate that, under some mild conditions on the sender's and receiver's preferences, the sender generically chooses at least partial disclosure over non-disclosure. Similarly, in our model, the non-disclosure conditions are stringent.

Miura's graph procedure plays a similar role as an edge between two nodes (states) in our graph procedure—it captures a motive for manipulative non-disclosure.

2.3 Model

We consider the standard Bayesian persuasion framework: a sender (S, she) designs and commits to an information structure (a Blackwell experiment) about an unknown state of the world $\omega \in \Omega$ to influence the action $a \in A$ of a receiver (R, he). The state space is finite, $\Omega \subset \mathbb{R}$, $|\Omega| = n$, and the action space is continuous, $A = \mathbb{R}$. The sender and the receiver have a common prior $p_0 \in \Delta(\Omega)$. They have the following preferences:

$$u_S = -(a - \omega)^2,$$
$$u_R = -(a - \rho(\omega))^2,$$

where $\rho : \Omega \to \mathbb{R}$ is arbitrary. Hence, state ω represents the preferred action of the sender and $\rho(\omega)$ the preferred action of the receiver.[6]

As is standard, the sender can be seen equivalently as choosing a Bayes-plausible distribution over posteriors, which we refer to as *signal*: $\pi \in \Delta(\Delta(\Omega))$ such that

$$\sum_{p \in supp(\pi)} \pi(p)p(\omega) = p_0(\omega) \ \forall \omega \in \Omega.$$[7]

The timing is as follows: the sender chooses a signal π, a posterior belief p is drawn according to π, and the receiver takes an action a given the belief p. The solution concept is subgame perfect equilibrium. The receiver's optimal action given a posterior belief p is $a(p) = \mathrm{E}_p[\rho(\omega)]$. Hence, using backward induction, the game reduces to the

[6] This model can be seen as a reduced form of a model in which the state of the world is two-dimensional, $y = (\omega_S, \omega_R)$, and the sender can design the experiment only about the dimension that is relevant for her, ω_S. The receiver then forms expectations about his relevant dimension, ω_R, using a common prior $p_0 \in \Delta(\Omega^2)$, so $\rho(\omega_S) = \mathrm{E}_{p_0}[\omega_R|\omega_S]$. This formulation maps better to the example with a politician and his advisor presented in the Introduction.

[7] Kamenica and Gentzkow (2011) show that there exists an optimal π such that $|supp(\pi)| \leq \min\{|\Omega|, |A|\}$. Hence, we restrict our search for the optimal signal only to signals satisfying $|supp(\pi)| \leq n$.

following problem of the sender:

$$\max_{\pi \in \Delta(\Delta(\Omega))} -\mathrm{E}_\pi \left[\mathrm{E}_p \left[(\mathrm{E}_p \left[\rho(\omega) \right] - \omega)^2 \right] \right] \text{ s.t. } \sum_{p \in supp(\pi)} \pi(p) p = p_0, \tag{2.1}$$

where $\mathrm{E}_\pi [\cdot]$ is the expectation over posteriors with respect to π and $\mathrm{E}_p [\cdot]$ is the expectation over states with respect to p.

2.4 General results about the optimal signal

In this section, we present general results about the optimal signal, and combine them in the next section to construct the procedure that allows us to discover which states are "pooled" together in the optimal signal.

To better understand how the sender chooses the signal, we start by inspecting the trade-off she faces. We can rewrite the objective function from her problem (2.1) as

$$\mathrm{var}_\pi \left(\mathrm{E}_p \left[\omega \right] \right) - \mathrm{E}_\pi \left[(\mathrm{E}_p \left[\omega - \rho(\omega) \right])^2 \right]. \tag{2.2}$$

The first term captures the benefit of a more informative (in the sense of Blackwell) π—this motive alone pushes her to reveal all states perfectly.[8] On the other hand, the second term captures the "cost" of revealed misalignment—this motive pushes her to "pool" some states to hide the largest misalignment. Hence, the sender prefers to reveal the most information so that the action is well adapted to the state. However, since she does not control the action directly, she wants to exploit the form of misalignment captured by ρ to manipulate the action of the receiver.

We can notice that the intercept of ρ does not play a role for the optimal signal. Formally, consider any function ρ and take $\rho' = b + \rho$ for some arbitrary constant $b \in \mathbb{R}$. The sender's objective function

$$\mathrm{var}_\pi \left(\mathrm{E}_p \left[\omega \right] \right) - \mathrm{E}_\pi \left[(\mathrm{E}_p \left[\omega - \rho'(\omega) \right])^2 \right]$$

[8] To illustrate this point, imagine an interior prior p_0, a signal π^1 with only interior beliefs, and a signal π^2 similar to π^1, but with more extreme beliefs: $p_k^2 = p_k^1 + \varepsilon(p_k^1 - p_0)\ \forall k$, for some small enough $\varepsilon > 0$. Then, $\mathrm{var}_{\pi^2} \left(\mathrm{E}_p \left[\omega \right] \right) = (1+\varepsilon)^2 \mathrm{var}_{\pi^1} \left(\mathrm{E}_p \left[\omega \right] \right) > \mathrm{var}_{\pi^1} \left(\mathrm{E}_p \left[\omega \right] \right)$.

can be rewritten in the form

$$\mathrm{var}_{\pi}\left(\mathrm{E}_{p}\left[\omega\right]\right)-\mathrm{E}_{\pi}\left[\left(\mathrm{E}_{p}\left[\omega-\rho(\omega)\right]\right)^{2}\right]-2b\mathrm{E}_{p_{0}}\left[\omega-\rho(\omega)\right]+b^{2}. \tag{2.3}$$

The last two terms in (2.3) do not depend on π, so the optimal signals under ρ and ρ' coincide. Hence, a state-independent bias b (no matter how large) does not affect the optimal signal.[9] Intuitively, the state-independent bias acts as a sunk cost for the sender. She cannot hide it by any manipulation of the signal because it is perfectly known ex ante.

It follows from the irrelevance of the intercept of ρ that what matters for the optimal signal is the overall shape of ρ, not agreement in particular states. In particular, perfect agreement between the sender and the receiver about the preferred action in a state of the world does not suffice for disclosure of that state. For example, consider two states $\omega_1 < \omega_2$, $\rho(\omega_1) = \omega_1$, $\rho(\omega_2) = 2\omega_1 - \omega_2$. Even though the sender and the receiver perfectly agree about the preferred action in ω_1, they substantially disagree in ω_2. It will be evident from the results in this section that full disclosure of the "perfect-agreement state" ω_1 is not optimal. Intuitively, due to the Bayesian consistency constraint, full disclosure of ω_1 would limit the opportunity to moderate the substantial disagreement in ω_2.[10]

2.4.1 Characterization of non-disclosure

In this subsection, we characterize the situation in which the sender does not benefit from revealing any information to the receiver.

Proposition 2.1. *The sender never (i.e., for any prior) benefits from providing any information if and only if ρ is linear with the slope from $(-\infty, 0] \cup [2, +\infty)$.*

Proof. The proof is in Appendix 2.A. It identifies the conditions for concavity of the expected utility of the sender as a function of the induced posterior by the principal-minor test of the Hessian matrix of this function. □

[9] We can contrast this feature with cheap talk (Crawford and Sobel 1982) in which the value of b matters for the informativeness of the equilibrium communication.

[10] In fact, Proposition 2.1 will imply that it is optimal not to disclose anything in this example.

Surprisingly, it is relatively easy to introduce some information revelation in our setting: it is sufficient to have a nonlinearity in ρ. The intuition for this generic taste for information revelation is that information has high value for the sender who wants to match the state of the world. The cases of optimal non-disclosure identified in Proposition 2.1 are intuitive too: (i) *misalignment in order*, i.e., when the sender and the receiver disagree about the order of the bliss actions (slope of ρ negative) or (ii) *misalignment in magnitude*, i.e., when they agree about the order, but the receiver overreacts relative to the sender (slope of ρ greater than two).

The non-disclosure characterized in Proposition 2.1 is never uniquely optimal for $n \geq 3$. To resolve such cases of indifference, we make the following assumption.

Assumption 2.1. Under indifference, the sender chooses not to disclose the states.

This assumption can be justified by the sender's interest in saving effort on communication when it is not needed. Technically, it greatly simplifies the analysis. Substantively, it leads us to identify the least informative signal in the indifference set of the sender. In Appendix 2.B, we analyze the structure of our problem that gives rise to the cases of indifference, and discuss the role of Assumption 2.1 as opposed to other selection criteria.

2.4.2 Full disclosure

In the next proposition, we provide a sufficient condition for full disclosure of the state of the world.

Proposition 2.2. *If ρ is linear with a slope in $[0, 2]$, full revelation of the state is always optimal (i.e., for any prior).*

Proof. The proof is in Appendix 2.A. It mostly follows from the proof of Proposition 2.1. □

For general n, Proposition 2.2 provides only a sufficient condition for full disclosure, but for $n = 2$ we can provide a full characterization. This special case is a cornerstone of our analysis of the case with general n.

Lemma 2.1. *For $n = 2$, the sender strictly prefers full revelation if and only if the slope of ρ is in $(0, 2)$. The sender is indifferent between any feasible signals if and only if the*

slope of ρ is either zero or two. The sender strictly prefers no revelation if and only if the slope of ρ is in $(-\infty, 0) \cup (2, \infty)$.

Proof. The proof is in Appendix 2.A. □

2.4.3 "Extremization"—non-existence of an interior posterior

After analyzing the conditions for extreme signals (non-disclosure and full disclosure), we look at signals that reveal some but not all information. The following proposition provides the key result enabling further analysis.

Proposition 2.3 (Extremization)**.** *Suppose non-disclosure is not optimal. Then, it is never optimal to induce an interior posterior.*

Proof. The proof is in Appendix 2.A. It is constructed by contradiction with the optimality of the signal, based on an improvement by splitting one of its posteriors. We call this result "extremization" because it leads us from the interior of the simplex to its extreme (boundary) subsimplexes as illustrated in Figure 2.2. This argument is independent of the prior, which is driven by the quadratic-utility assumption. □

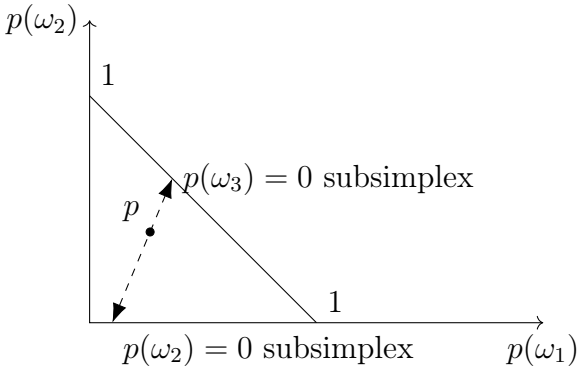


Figure 2.2: Illustration of the extremization result from Proposition 2.3 for three states $\omega_1, \omega_2, \omega_3$: if non-disclosure is not optimal, then for any interior posterior p, there exists a direction (dashed line) along which it is strictly beneficial to split p all the way to the boundary subsimplexes (in the figure, subsimplexes corresponding to $p(\omega_2) = 0$ and $p(\omega_3) = 0$); this insight can be used again on the boundary subsimplexes

We can apply Proposition 2.3 iteratively to eliminate the areas of posteriors that will not appear in the optimal signal. This sharpens the idea about the structure of the optimal

signal, which is our main interest, and simplifies the search for it. We use this idea in the next section.

2.5 State-pooling structure of the optimal signal

In this section, we go beyond the extreme cases of full disclosure and non-disclosure and study how preference misalignment, captured by ρ, affects a qualitative property of the optimal signal that we call state pooling. We define the state-pooling structure of a signal and present an illustrative procedure for its discovery that builds on the general results from Section 2.4.

2.5.1 Definitions

Definition 2.1. We say that states $\omega_{k_1}, \ldots, \omega_{k_m}$, for some $k_1, \ldots, k_m \in \{1, \ldots, n\}$, are *pooled* together (or form a *pool of states*) under signal π if the set $M = \{\omega_{k_1}, \ldots, \omega_{k_m}\}$ satisfies

$$\exists p \in supp(\pi) : supp(p) = M \;\&\; \forall p' \in supp(\pi) \text{ s.t. } p' \neq p : M \nsubseteq supp(p'),$$

where $supp(\cdot)$ denotes support. The set of all pools of states that signal π induces is called the *state-pooling structure* of signal π.

In intuitive terms, we say that states $\omega_{k_1}, \ldots, \omega_{k_m}$ are pooled together under signal π if π reveals whether the event $\{\omega_{k_1}, \ldots, \omega_{k_m}\}$ occurred. A pool is the largest set of states that appears in a support of a posterior. There may be several largest pools, e.g., $\{\omega_1\}$ and $\{\omega_2, \omega_3\}$. Notice that if there exists an interior posterior under π, then the pooling structure consists only of the set of all states. Hence, the study of pooling structures might not be interesting in general setups. In our setup, however, the study of pooling structures brings interesting insights thanks to the extremization result (Proposition 2.3).

The state-pooling structure of a signal can be captured *graphically* by representing each state of the world by a node and each pool by highlighting the corresponding set of nodes; an example is presented in Figure 2.3.

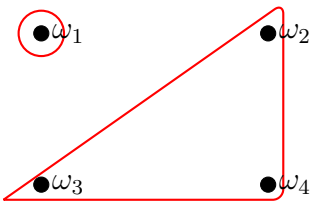


Figure 2.3: Example of a graphical representation of the state-pooling structure when $n = 4$ and the signal induces posteriors supported on $\{\omega_1\}$ and $\{\omega_2, \omega_3, \omega_4\}$

In the next subsection, we propose a procedure that aims to find the state-pooling structure of the optimal signal for a given form of preference misalignment captured by ρ. This procedure can easily be represented graphically; its desired output is a graphical representation of the state-pooling structure of the type depicted in Figure 2.3, i.e., nodes representing states and highlighted pools. However, the proposed procedure may not identify the state-pooling structure of the optimal signal completely in some cases, but may offer only candidates for optimal pools. Nevertheless, we can often identify which of the candidate pools are certainly a part of the optimal state-pooling structure. Hence, we introduce two types of highlighting in the procedure—*dashed* (highlighting candidate pools) and *full* (highlighting pools certainly belonging to the optimal state-pooling structure). Naturally, highlighting in full is superior to highlighting in dashed because it expresses certainty.

An important working component of the graphical procedure are the *edges* between pairs of nodes—they represent a pooling tendency of the corresponding states. We will see that this pooling tendency is driven by the slope of ρ between pairs of corresponding states; we denote the *slope of* ρ *between states* ω_i *and* ω_j by

$$s_{ij} = \frac{\rho(\omega_j) - \rho(\omega_i)}{\omega_j - \omega_i}. \tag{2.4}$$

This object represents an index of misalignment between the receiver (the numerator) and the sender (the denominator).[11]

A subroutine of our procedure relates to the well-known problem from computer science called the *clique problem*. Thus, we borrow a few notions from graph theory.

[11] A similar object plays an important role for the pooling structure (of types) in Hummel, Morgan, and Stocken (2018).

Definition 2.2. Let $G = (V, E)$ be an undirected graph (with V denoting the set of nodes and E denoting the set of edges). We call a subset of nodes $C \subseteq V$ a *clique* if the subgraph of G induced by C is complete (i.e., the nodes in C are fully connected). A *clique* C is called *maximal* if there does not exist another clique strictly larger than C (in the sense of inclusion).

The version of the clique problem that we are interested in is finding all maximal cliques in an undirected graph. Systematic inspection of all subsets of nodes or the *Bron–Kerbosch algorithm* can be used to solve this problem.

2.5.2 Procedure for discovery of the state-pooling structure of the optimal signal

We present a procedure that inspects the form of misalignment function ρ and reflects its implications for the state-pooling structure of the optimal signal on a graph. The output are pools highlighted in full (which are certainly present in the state-pooling structure of the optimal signal) and candidate pools highlighted in dashed (which may be present in the state-pooling structure of the optimal signal). We present an example of the output of this procedure at the end of this subsection and a step-by-step illustration of the procedure leading to this output in Appendix 2.C.

Procedure for discovery of the state-pooling structure of the optimal signal:
Input: Set of states Ω ($|\Omega| = n$) and preference-misalignment function $\rho : \Omega \to \mathbb{R}$.

1. Create a fully connected graph on n nodes where node i corresponds to state ω_i.
2. Eliminate all edges ij such that the slope of ρ on $\omega_i < \omega_j$, s_{ij}, is in $(0, 2)$.
3. Highlight in full each isolated node (i.e., a node with no edges leading to any other node) as a singleton pool.
4. Among the remaining (i.e., non-isolated) nodes, list all maximal cliques.
5. For each maximal clique C:
   for k from $|C|$ to 2:
   for all subsets $M \subseteq C$ such that $|M| = k$:
   - If M was ever inspected before, do nothing and continue iteration.

- If M is a subset of a highlighted set of nodes, do nothing and continue iteration.
- Otherwise, apply the non-disclosure test to the inspected pool M: Is ρ linear with slope in $(-\infty, 0] \cup [2, \infty)$ on the states corresponding to the nodes in M?
  - If yes, highlight pool M in dashed on output and continue iteration.
  - If no, denote M as inspected and continue iteration.

6. If any node belongs only to one highlighted pool (in dashed), highlight the corresponding pool in full (if not already highlighted in full).

An example of the output produced by this procedure appears in the right panel of Figure 2.4; an example of function ρ leading to this output is depicted in the left panel.[12] State 1 is isolated because the sender and the receiver agree on its position relative to other states both in order and in magnitude, so there is no reason for the sender to leverage this state for manipulation of beliefs. States 2, 3, and 4 are pooled together (they pass the non-disclosure test) because the sender tries to moderate the action of the receiver, who would overreact in these states (disagreement about magnitude). States 3 and 5 may be pooled together (disagreement about order) and 4 and 5 may also be pooled together (disagreement about order), but states 3, 4, and 5 are not pooled together even though they form a maximal clique (because they do not pass the non-disclosure test)—the sender prefers to exploit some variation in this collection of nodes. Hence, the optimal signal will induce posterior $p_1 = \delta_1$ and posterior p_2 supported on 2, 3, and 4. Moreover, it will induce at least one of the posteriors p_3 or p_4 supported on 3 and 5 or 4 and 5, respectively.

2.5.3 Discussion of the procedure

The idea underlying our proposed procedure is the iterative application of Proposition 2.1 and Proposition 2.3, which we call a *top-down approach*. Starting from the full $(n-1)$-dimensional simplex,[13] we can check whether non-disclosure is optimal using

[12] A step-by-step illustration of the procedure leading to this output appears in Appendix 2.C.

[13] We start from the $(n-1)$-dimensional simplex because $p_n = 1 - p_1 - \cdots - p_{n-1}$.

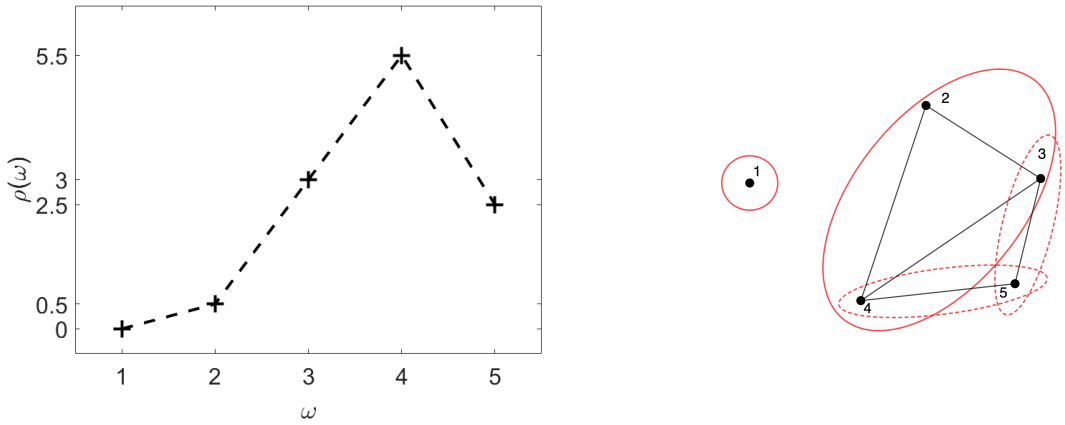


Figure 2.4: Output of the graph procedure (right panel) for function ρ on the left panel: 1 is isolated; 2, 3, and 4 are pooled together (they pass the non-disclosure test); 3 and 5 may be pooled together; 4 and 5 may be pooled together; 3, 4, and 5 are not pooled together (they do not pass the non-disclosure test)

Proposition 2.1. If it is optimal, the sender chooses a completely uninformative signal. If it is not, Proposition 2.3 suggests that the optimal signal will induce posteriors on the boundary of the $(n-1)$-dimensional simplex. Hence, we focus on each of the $(n-2)$-dimensional boundary simplexes and apply the same test. Specifically, by restricting the sender's expected utility (as a function of the posterior) on a particular $(n-2)$-dimensional simplex, we use Proposition 2.1 to check if non-disclosure is optimal there:

- If it is optimal, then the sender cannot benefit from splitting the pool of states corresponding to the vertices of the inspected $(n-2)$-dimensional simplex. However, the sender might not want to choose this pool of states at all, so this pool of states constitutes only a candidate pool for the optimal signal.[14]

- If it is not optimal, then by Proposition 2.3 we eliminate all interior points from the inspected $(n-2)$-dimensional simplex and restrict our focus to its $(n-3)$-dimensional boundary simplexes; for each of them, we repeat the same steps.

Along the path from the full $(n-1)$-dimensional simplex to lower-dimensional simplexes due to elimination of "interior" posteriors outlined in the second bullet point, we move closer to the trivial case of 1-dimensional simplexes where we apply Lemma 2.1.

Our procedure relies on this top-down approach in Step 5. However, compared to the top-

[14]Here, we also use Assumption 2.1. This simplifies the analysis because we do not need to keep track of all equivalent splits.

down approach, the procedure starts with a simplification of the problem by identifying the only relevant subsets of nodes for this inspection—the maximal cliques (Steps 2 and 4). This step is justified by the fact that the necessary condition for optimality of non-disclosure on a simplex is optimality of non-disclosure on its boundary simplexes, which follows easily from Proposition 2.1. Hence, if we have a given collection of nodes with some pair of nodes in it that is not pooled, this whole collection of nodes cannot form a pool.

In Steps 3 and 6 of the procedure, we exploit Bayesian consistency (and the interior prior). In particular, the structure of the graph obtained after Step 2 is informative about the state-pooling structure by itself: any isolated node represents a state that is fully disclosed. In Step 5, we can identify only candidates for optimal pools, but, in Step 6, Bayesian consistency can help us to determine which of them will be certainly a part of the optimal pooling structure.

Note that we have not mentioned the prior in our identification of the optimal pooling structure. This prior-independence of our procedure relies on a feature of the quadratic setting: constant convexity/concavity structure in all points. However, even in the quadratic setting, the pooling structure of the optimal signal itself is not always prior-independent. This feature imposes a limit on how far we can go with our simple prior-independent procedure in identifying the full pooling structure of the optimal signal. In some cases, we also need to incorporate the prior into our analysis at the end of the procedure (see Section 2.6 for examples).

2.6 Characterization of the state-pooling structure for $n = 3$

In this section, we use the above procedure to characterize the state-pooling structure of the optimal signal in the simplest interesting case of three states (the case of two states is trivial and is fully characterized in Lemma 2.1). We describe the state-pooling structure for all possible cases of the form of ρ, which we capture through s_{12}, s_{23}, and s_{13}. For clarity of exposition, we divide the cases into five classes (i)-(v) based on the features of the resulting state-pooling structure and the role of the prior. Class (i) corresponds to full disclosure, class (ii) corresponds to signals that fully disclose one of the states, classes

(iii) and (iv) correspond to signals that reveal some information without fully revealing any of the states, and class (v) corresponds to non-disclosure. Within a given class, we use letters to distinguish between particular state-pooling structures.

Theorem 2.1. *Assume that there are three states of the world,* $\Omega = \{\omega_1, \omega_2, \omega_3\}$. *Depending on the form of* ρ, *as pinned down by* s_{12}, s_{23}, *and* s_{13}, *the state-pooling structure of the optimal signal is as follows:*

	s_{12}	s_{23}	s_{13}	state-pooling structure
i	$\in (0,2)$	$\in (0,2)$	$\in (0,2)$	$\{\{\omega_1\},\{\omega_2\},\{\omega_3\}\}$
ii.a	$\in (0,2)$	$\notin (0,2)$	$\in (0,2)$	$\{\{\omega_1\},\{\omega_2,\omega_3\}\}$
ii.b	$\notin (0,2)$	$\in (0,2)$	$\in (0,2)$	$\{\{\omega_3\},\{\omega_1,\omega_2\}\}$
iii.a	$\notin (0,2)$	$\in (0,2)$	$\notin (0,2)$	$\{\{\omega_1,\omega_2\},\{\omega_1,\omega_3\}\}$
iii.b	$\in (0,2)$	$\notin (0,2)$	$\notin (0,2)$	$\{\{\omega_2,\omega_3\},\{\omega_1,\omega_3\}\}$
iii.c	$\notin (0,2)$	$\notin (0,2)$	$\in (0,2)$	$\{\{\omega_1,\omega_2\},\{\omega_2,\omega_3\}\}$
iv[15]	$\notin (0,2)$	$\notin (0,2)$	$\notin (0,2)$	depending on s_{12}, s_{23}, s_{13}, and prior, either (iii.a), (iii.b), or (iii.c) pooling
v	$s_{12} = s_{23} = s_{13} = s \notin (0,2)$			$\{\{\omega_1,\omega_2,\omega_3\}\}$

Proof. The proof is in Appendix 2.A. □

The observed state-pooling structures emerge from the interaction of the two main forces that drive the sender's choice. On the one hand, the sender wants to disclose the states so that the induced receiver's actions vary sufficiently with the state of the world. On the other hand, she wants to pool the states together to dampen that variation if there is a severe misalignment in either order or magnitude in some pairs of states. The slope of ρ for states ω_i and ω_j, s_{ij} $(i, j \in \{1,2,3\},\ i \neq j)$, serves as an index that can capture the misalignment in either order or magnitude in that pair of states.

In case (i), there is no severe preference misalignment in either pair of states, so the sender fully discloses each state. In case (ii.a), s_{23} captures a severe preference misalignment in the pair of states ω_2, ω_3, so the sender pools these states together to conceal the misalignment but reveals state ω_1 to maximize the informativeness of the signal. In case (iii.a), s_{12}

[15] $s_{12} = s_{23} = s_{13} = s \notin (0,2)$ corresponds to non-disclosure, so we exclude this combination from case (iv) and denote it as a separate case (v). See Appendix 2.A for details on the choice from (iii.a), (iii.b), and (iii.c).

and s_{13} capture a severe preference misalignment in two pairs of states, so the sender pools the respective pairs together but still reveals some information: $\{\{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}\}$. In case (iv), there is a misalignment in each of the three pairs of states and the optimal state-pooling structure is sensitive to the prior and to the relation between the slopes of ρ.

A notable feature of the state-pooling structure of the optimal signal under $n = 3$ is that the sender never chooses to fully disclose the middle state of the world ω_2 and pool ω_1 and ω_3 together. For that to be the case, it would need to hold $s_{13} \notin (0, 2)$, $s_{12} \in (0, 2)$, and $s_{23} \in (0, 2)$, which cannot happen.[16] The intuition is that full disclosure of ω_2 and pooling of ω_1 and ω_3 is not in line with the sender's preference for maximizing the variance of the induced posterior beliefs. A potentially better way to leverage state ω_2 is to form two pools $\{\omega_2, \omega_1\}$ and $\{\omega_2, \omega_3\}$ because it can induce relatively more variation in the receiver's actions.

2.7 Conclusion

We consider a Bayesian persuasion model in which both the sender and the receiver have state-dependent preferred actions. We specialize to a quadratic-utility setting to simplify the otherwise nontrivial problem of characterizing the optimal signal. In this framework, we make the trade-off that drives the sender's choice of the signal transparent: on the one hand, the sender wants to reveal information to adapt the action to the state of the world; on the other hand, she wants to hide information to conceal the misalignment between her and the receiver.

We focus on characterization of the state-pooling structure of the optimal signal. In particular, we link the form of misalignment between the sender and the receiver in their preferred (state-dependent) actions to the state-pooling structure of the sender's optimal signal. To achieve this goal, we propose an illustrative graphical procedure for finding the sets of states that are pooled together in the supports of posteriors of the optimal signal.

[16] Note that $s_{13} = \frac{\rho(\omega_3) - \rho(\omega_1)}{\omega_3 - \omega_1} = \frac{1}{(\omega_3 - \omega_2) + (\omega_2 - \omega_1)} \left(s_{23} (\omega_3 - \omega_2) + s_{12} (\omega_2 - \omega_1)\right)$ and $(0, 2)$ is a convex set.

Our model naturally suits the analysis of influence in political economy. The sender's and receiver's (state-dependent) single-peaked preferences over the continuous action space are consistent with ideology-based preferences over a continuous set of policy alternatives. That set could represent potential allocations of a resource such as the amount of budget spending on a public good. Thus, our framework can capture an arbitrary form of ideological disagreement between a lobbyist and a policymaker regarding the preferred state-dependent policy and yield predictions about the structure of the lobbyist's chosen information disclosure.

Our analysis motivates a number of directions for further research. First, further investigation and economic interpretation of particular state-pooling patterns that emerge when there are more than three states of the world might be of interest. Second, more progress could be made on analyzing state-pooling patterns that may emerge under loss functions of a more general form.

2.A Technical details and proofs

2.A.1 The structure of the sender's problem

We are interested in the solution of the sender's problem

$$\max_{\pi\in\Delta(\Delta(\Omega))} -\mathrm{E}_\pi\left[\mathrm{E}_p\left[\left(\mathrm{E}_p\left[\rho(\omega)\right]-\omega\right)^2\right]\right] \text{ s.t. } \sum_{p\in supp(\pi)}\pi(p)p=p_0.$$

We can rewrite the objective function as

$$-\mathrm{E}_\pi\left[\mathrm{E}_p\left[\rho(\omega)\right]^2-2\mathrm{E}_p\left[\rho(\omega)\right]\mathrm{E}_p\left[\omega\right]+\mathrm{E}_p\left[\omega^2\right]\right].$$

Using the Bayesian consistency condition $\sum_p \pi(p)p=p_0$, we can see that the last term becomes

$$-\mathrm{E}_{p_0}\left[\omega^2\right].$$

Therefore, the solution to the problem above is the same as the solution to the problem

$$\max_{\pi\in\Delta(\Delta(\Omega))} \mathrm{E}_\pi\left[\mathrm{E}_p\left[\rho(\omega)\right]\left(2\mathrm{E}_p\left[\omega\right]-\mathrm{E}_p\left[\rho(\omega)\right]\right)\right] \text{ s.t. } \sum_{p\in supp(\pi)}\pi(p)p=p_0.$$

A general approach to solving this problem is concavification of the function

$$g(p)=\mathrm{E}_p\left[\rho(\omega)\right]\left(2\mathrm{E}_p\left[\omega\right]-\mathrm{E}_p\left[\rho(\omega)\right]\right). \tag{2.5}$$

We use the parametrization $g(p)=g(p_1,p_2,\ldots,p_{n-1})$, where $p_n=1-p_1-\cdots-p_{n-1}$. We collect the free variables in the vector

$$\bar{p}=(p_1,\ldots,p_{n-1})'.$$

We also denote

$$\bar{\rho}=(\rho(\omega_1)-\rho(\omega_n),\ldots,\rho(\omega_{n-1})-\rho(\omega_n))',$$
$$\bar{\omega}=(\omega_1-\omega_n,\ldots,\omega_{n-1}-\omega_n)'.$$

With this notation, we can write

$$g(\bar{p}) = \bar{p}' \underbrace{[2\bar{\rho}\bar{\omega}' - \bar{\rho}\bar{\rho}']}_{G} \bar{p} + [2\omega_n\bar{\rho}' - \rho_n\bar{\rho}' + 2\rho_n\bar{\omega}' - \rho_n\bar{\rho}']\bar{p} + 2\rho_n\omega_n - \rho_n^2.$$

Hence, the curvature of g is driven by matrix G because the Hessian matrix is

$$H = G + G'.^{17}$$

The ij element ($i, j \in \{1, \ldots, n-1\}$) of H is

$$H_{ij} = \frac{\partial^2 g(p)}{\partial p_i \partial p_j} = 2\{[\rho(\omega_i) - \rho(\omega_n)](\omega_j - \omega_n) - [\rho(\omega_i) - \rho(\omega_n)][\rho(\omega_j) - \rho(\omega_n)] + [\rho(\omega_j) - \rho(\omega_n)](\omega_i - \omega_n)\}.$$

This special structure of the problem implies that general submatrices of order 3 (for $n \geq 4$) of the Hessian matrix H have zero determinants.[18] Hence, by the Laplace expansion of determinants, all submatrices of order $k \geq 3$ have zero determinants. We can deduce from this observation, using the fact that the determinant rank of a matrix is equal to the column/row rank of the matrix,[19] that H has at most two non-zero eigenvalues. Therefore, there are at least $n-3$ orthogonal directions (in space $\mathbb{R}^{n-1} \ni \bar{p}$) that span the space along which g is linear, and at most two orthogonal directions that span the space (orthogonal to the space spanned by the linear directions) on which g has a less trivial shape.

---

[17] We can also rewrite g as a linear-quadratic form

$$g(\bar{p}) = \frac{1}{2}\bar{p}' H \bar{p} + [2\omega_n\bar{\rho}' - \rho_n\bar{\rho}' + 2\rho_n\bar{\omega}' - \rho_n\bar{\rho}']\bar{p} + 2\rho_n\omega_n - \rho_n^2.$$

[18] Consider a submatrix corresponding to rows i_1, i_2, i_3 and columns j_1, j_2, j_3. Then column j_3 of this submatrix is a linear combination of columns j_1 and j_2 with the following vector of coefficients

$$\begin{pmatrix} -\frac{\rho(\omega_{i_2})(\omega_{j_3}-\omega_n)-\rho(\omega_{i_3})(\omega_{j_2}-\omega_n)+\rho(\omega_n)(\omega_{j_2}-\omega_{j_3})}{\rho(\omega_{i_1})(\omega_{j_2}-\omega_n)-\rho(\omega_{i_2})(\omega_{j_1}-\omega_n)+\rho(\omega_n)(\omega_{j_1}-\omega_{j_2})} \\ \frac{\rho(\omega_{i_1})(\omega_{j_3}-\omega_n)-\rho(\omega_{i_3})(\omega_{j_1}-\omega_n)+\rho(\omega_n)(\omega_{j_1}-\omega_{j_3})}{\rho(\omega_{i_1})(\omega_{j_2}-\omega_n)-\rho(\omega_{i_2})(\omega_{j_1}-\omega_n)+\rho(\omega_n)(\omega_{j_1}-\omega_{j_2})} \end{pmatrix}.$$

[19] The *determinant rank* of H is the size k of the largest $k \times k$ submatrix with a non-zero determinant. The *column/row rank* of H is the dimension of the space spanned by the columns/rows of H. It is straightforward to show that these ranks are equal.

2.A.2 Proofs

Proof of Proposition 2.1. The sender does not benefit from providing any information if and only if g is concave.[20] g is concave if and only if its Hessian matrix is negative semidefinite, which can be checked with the test on its principal minors.

Suppose $n \geq 3$ (the case $n = 2$ is covered separately in Lemma 2.1). Let Δ_k be a principal minor of order k of the Hessian matrix of g. Since $\Delta_k = 0$ for $k \geq 3$ (see the discussion above), a necessary and sufficient condition for g to be concave is $\Delta_1 \leq 0$ and $\Delta_2 \geq 0$ for all Δ_1, Δ_2.

Let Δ_1^i be the first-order principal minor obtained from row (column) i:

$$\Delta_1^i = 2\left(\rho\left(\omega_i\right) - \rho\left(\omega_n\right)\right)\left(2\left(\omega_i - \omega_n\right) - \left(\rho\left(\omega_i\right) - \rho\left(\omega_n\right)\right)\right). \tag{2.6}$$

Let Δ_2^{ij} be the second-order principal minor obtained from rows (columns) i and j:

$$\Delta_2^{ij} = -4[(\rho(\omega_i) - \rho(\omega_j))(\omega_j - \omega_n) - (\rho(\omega_j) - \rho(\omega_n))(\omega_i - \omega_j)]^2.$$

We can see that $\Delta_2^{ij} \leq 0$. Hence, g is concave or convex only if $\Delta_2 = 0$ for all Δ_2. This condition yields a system of $\frac{(n-1)(n-2)}{2}$ equations

$$\Delta_2^{ij} = 0, \;\; i, j \in \{1, ..., n-1\}, \;\; i \neq j. \tag{2.7}$$

Under the natural assumption that $\omega_1 < \cdots < \omega_n$ (which is without loss of generality), we obtain from $\Delta_2^{ij} = 0$

$$\frac{\rho\left(\omega_j\right) - \rho\left(\omega_i\right)}{\omega_j - \omega_i} = \frac{\rho\left(\omega_n\right) - \rho\left(\omega_j\right)}{\omega_n - \omega_j} \tag{2.8}$$

or, equivalently,

$$\frac{\rho\left(\omega_j\right) - \rho\left(\omega_i\right)}{\omega_j - \omega_i} = \frac{\rho\left(\omega_n\right) - \rho\left(\omega_i\right)}{\omega_n - \omega_i}. \tag{2.9}$$

Therefore, the system of equations (2.7) gives rise to $\frac{(n-1)(n-2)}{2}$ slope equality conditions.

[20] The "if" part follows directly from the definition of concavity. The "only if" part would also follow directly from the definition of concavity if the sender did not benefit from providing any information for every prior. But if the sender does not benefit from providing any information only in one prior, because g is a linear-quadratic form, this property extends to all priors.

From (2.8) and (2.9), we have

$$\begin{aligned}
j=n-1, i=n-2: \frac{\rho(\omega_n)-\rho(\omega_{n-1})}{\omega_n-\omega_{n-1}} &= \frac{\rho(\omega_{n-1})-\rho(\omega_{n-2})}{\omega_{n-1}-\omega_{n-2}} = \frac{\rho(\omega_n)-\rho(\omega_{n-2})}{\omega_n-\omega_{n-2}},\\
j=n-2, i=n-3: \frac{\rho(\omega_n)-\rho(\omega_{n-2})}{\omega_n-\omega_{n-2}} &= \frac{\rho(\omega_{n-2})-\rho(\omega_{n-3})}{\omega_{n-2}-\omega_{n-3}} = \frac{\rho(\omega_n)-\rho(\omega_{n-3})}{\omega_n-\omega_{n-3}},\\
&\vdots\\
j=2, i=1: \frac{\rho(\omega_n)-\rho(\omega_2)}{\omega_n-\omega_2} &= \frac{\rho(\omega_2)-\rho(\omega_1)}{\omega_2-\omega_1} = \frac{\rho(\omega_n)-\rho(\omega_1)}{\omega_n-\omega_1}.
\end{aligned}$$

Hence, system (2.7) is equivalent to a linearity of ρ:

$$s := \frac{\rho(\omega_2)-\rho(\omega_1)}{\omega_2-\omega_1} = \frac{\rho(\omega_3)-\rho(\omega_2)}{\omega_3-\omega_2} = \cdots = \frac{\rho(\omega_n)-\rho(\omega_{n-1})}{\omega_n-\omega_{n-1}}.$$

Finally, given that $\Delta_2 = 0$ for all Δ_2 holds, one can establish whether g is concave or convex based on the sign of Δ_1. Inspecting the sign of (2.6) yields:

$$\Delta_1^i \geq 0 \iff (\rho(\omega_n)-\rho(\omega_i)\geq 0) \wedge \frac{\rho(\omega_n)-\rho(\omega_i)}{\omega_n-\omega_i} \leq 2 \iff 0 \leq s \leq 2.$$

The complement identifies the concavity slopes (including the borderline slopes $s \in \{0,2\}$). □

Proof of Proposition 2.2. This proposition is basically proven in the proof of Proposition 2.1, using the fact that g is convex if and only if $\Delta_1 \geq 0$ and $\Delta_2 \geq 0$ for all Δ_1, Δ_2. The only difference is that the convexity of g is only sufficient for optimality of full disclosure, but is not necessary (we can provide an example of optimal full disclosure with non-convex g). □

Proof of Lemma 2.1. For $n=2$, g is a quadratic function, so its second derivative completely characterizes its curvature, which completely characterizes the type of optimal signals. In particular, let $\omega_1 < \omega_2$. Then,

$$\frac{\partial^2 g(p_1)}{\partial p_1^2} = 2(\rho(\omega_1)-\rho(\omega_2))(2(\omega_1-\omega_2)-(\rho(\omega_1)-\rho(\omega_2))),$$

which is strictly positive if and only if the slope of ρ is in $(0,2)$ (strict convexity and full disclosure), strictly negative if and only if the slope of ρ is in $(-\infty,0)\cup(2,\infty)$ (strict

concavity and non-disclosure), and zero if and only if the slope of ρ is either zero or two (linearity and indifference). □

Proof of Proposition 2.3. Non-disclosure is optimal if and only if g is concave. Hence, if non-disclosure is not optimal, g is not concave. Therefore, g has to have a direction along which it is strictly convex. Since g is a linear-quadratic form, the existence of this convex direction holds at any point p and is global along that direction, i.e., for any belief parametrized by $p \in \mathbb{R}^{n-1}$, there exists a direction $v \in \mathbb{R}^{n-1}$ such that function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(\lambda) := g(p + \lambda v)$ is strictly convex.

Suppose (toward contradiction) that it is optimal to induce an interior posterior, i.e., there exists a posterior p in the support of the optimal signal π such that $p(\omega) > 0 \ \forall \omega$. Then, we can split p along a strictly convex direction to q_1 and q_2, i.e., there exists some $\lambda \in (0, 1)$ such that $p = \lambda q_1 + (1 - \lambda) q_2$. Then, π' formed from π by replacing p by q_1 with probability $\lambda \pi(p)$ and q_2 with probability $(1 - \lambda)\pi(p)$ is Bayes-plausible and it induces a strict improvement for the sender because, from strict convexity of g along the direction determined by q_1 and q_2,

$$\mathrm{E}_{\pi'}\left[g(p)\right] - \mathrm{E}_{\pi}\left[g(p)\right] = \pi(p)(\lambda g(q_1) + (1 - \lambda) g(q_2) - g(p)) > 0.$$

This is a contradiction with optimality of π. □

Proof of Proposition 2.1. We derive the state-pooling structure for the form of ρ for each case presented in the table of Proposition 2.1 using the graph procedure presented in Section 2.5.2.

Case (i). Since $s_{12}, s_{23}, s_{13} \in (0, 2)$, Step 2 of the procedure eliminates all edges, so each node is highlighted in full in Step 3. Thus immediately after Step 3, the procedure yields the state-pooling structure of the optimal signal $\{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\}$.

Case (ii.a). Since $s_{12}, s_{13} \in (0, 2)$ and $s_{23} \notin (0, 2)$, after Step 2 of the procedure, node 1 is isolated (thus, it is highlighted in full in Step 3) and there is an edge left between nodes 2 and 3. Since the pool $\{2, 3\}$ is a maximal clique (Step 4) and ρ is obviously linear with slope from $(-\infty, 0] \cup [2, \infty)$ on states ω_2 and ω_3, this pool is highlighted in dashed in Step 5. Finally, it is highlighted in full in Step 6 because nodes 2 and 3 belong only to this pool. Therefore, the state-pooling structure of the optimal signal is $\{\{\omega_1\}, \{\omega_2, \omega_3\}\}$.

Case (ii.b). Analogous to case (ii.a).

Case (iii.a). Since $s_{12}, s_{13} \notin (0,2)$ and $s_{23} \in (0,2)$, after Step 2 of the procedure, there are two edges left: one between nodes 1 and 2 and one between nodes 1 and 3. Since both pools $\{1,2\}$ and $\{1,3\}$ are maximal cliques (Step 4) and ρ is obviously linear with slope from $(-\infty,0]\cup[2,\infty)$ on states ω_1,ω_2 and ω_1,ω_3, respectively, these pools are highlighted in dashed in Step 5. Finally, they are highlighted in full in Step 6 because node 2 belongs only to pool $\{1,2\}$ and node 3 belongs only to pool $\{1,3\}$. Therefore, the state-pooling structure of the optimal signal is $\{\{\omega_1,\omega_2\},\{\omega_1,\omega_3\}\}$.

Case (iii.b). Analogous to case (iii.a).

Case (iii.c). Analogous to case (iii.a).

Case (iv). We assume that $s_{12} = s_{23} = s_{13} = s \notin (0,2)$ does not hold (this case is covered by case (v)). Thus, the graph procedure yields the candidate pools $\{\omega_1,\omega_2\}$, $\{\omega_2,\omega_3\}$, and $\{\omega_1,\omega_3\}$ (corresponding to the pools of nodes highlighted in dashed in the graph). To determine the optimal state-pooling structure given the set of candidate pools is non-trivial.

Denote the n-th directional derivative of a function $f:\mathbb{R}^2\to\mathbb{R}$ along a direction (a,b) by $D^n_{(a,b)}f$. Denote $p_1 \coloneqq \Pr(\omega_1)$ and $p_2 \coloneqq \Pr(\omega_2)$. From the proof of Proposition 2.1, the nonlinearity in ρ implies that there exists a direction (a,b) along which $g(p_1,p_2)$ (defined in (2.5)) is strictly convex. The set of all such directions is pinned down by the condition

$$D^2_{(a,b)}g(p) > 0,$$

which rewrites as (assuming $s_{13}\neq 0$ and $s_{23}\neq 0$; see below for the discussion of these cases)

$$\begin{gathered}
a^2(\rho(\omega_1)-\rho(\omega_3))[2(\omega_1-\omega_3)-(\rho(\omega_1)-\rho(\omega_3))]+\\
b^2(\rho(\omega_2)-\rho(\omega_3))[2(\omega_2-\omega_3)-(\rho(\omega_2)-\rho(\omega_3))]+\\
ab\tfrac{\rho(\omega_2)-\rho(\omega_3)}{\rho(\omega_1)-\rho(\omega_3)}(\rho(\omega_1)-\rho(\omega_3))[2(\omega_1-\omega_3)-(\rho(\omega_1)-\rho(\omega_3))]+\\
ab\tfrac{\rho(\omega_1)-\rho(\omega_3)}{\rho(\omega_2)-\rho(\omega_3)}(\rho(\omega_2)-\rho(\omega_3))[2(\omega_2-\omega_3)-(\rho(\omega_2)-\rho(\omega_3))]>0.
\end{gathered} \tag{2.10}$$

Next, $s_{13} \notin (0,2) \wedge s_{23} \notin (0,2)$ implies[21]

$$\begin{cases} (\rho(\omega_1) - \rho(\omega_3))\left[2(\omega_1 - \omega_3) - (\rho(\omega_1) - \rho(\omega_3))\right] & \leq 0, \\ (\rho(\omega_2) - \rho(\omega_3))\left[2(\omega_2 - \omega_3) - (\rho(\omega_2) - \rho(\omega_3))\right] & \leq 0. \end{cases} \tag{2.11}$$

We can see from (2.10) and (2.11) that if (a, b) is a direction along which g is strictly convex, both a and b have to be non-zero. Thus, we can normalize the direction (a, b) to $(\frac{a}{b}, 1)$ and denote $x := \frac{a}{b}$. Hence, the set of directions along which g is strictly convex is characterized by

$$\begin{gathered} x^2(\rho(\omega_1) - \rho(\omega_3))\left[2(\omega_1 - \omega_3) - (\rho(\omega_1) - \rho(\omega_3))\right] + \\ (\rho(\omega_2) - \rho(\omega_3))\left[2(\omega_2 - \omega_3) - (\rho(\omega_2) - \rho(\omega_3))\right] + \\ x\frac{\rho(\omega_2)-\rho(\omega_3)}{\rho(\omega_1)-\rho(\omega_3)}(\rho(\omega_1) - \rho(\omega_3))\left[2(\omega_1 - \omega_3) - (\rho(\omega_1) - \rho(\omega_3))\right] + \\ x\frac{\rho(\omega_1)-\rho(\omega_3)}{\rho(\omega_2)-\rho(\omega_3)}(\rho(\omega_2) - \rho(\omega_3))\left[2(\omega_2 - \omega_3) - (\rho(\omega_2) - \rho(\omega_3))\right] > 0. \end{gathered} \tag{2.12}$$

Inspecting (2.12) given (2.11), one observes that the first two terms in (2.12) are non-positive. Therefore, the sum of the last two terms must necessarily be strictly positive for any direction along which g is strictly convex. Further, if the third term is strictly negative, the fourth term is non-positive and vice versa. So, if either of the last two terms is strictly negative, their sum is also strictly negative. Equivalently, if their sum is non-negative, they both have to be non-negative. Moreover, if their sum is strictly positive, they cannot both be zero. But if any one of the last two terms in (2.12) is strictly positive, then by (2.11)

$$x\frac{\rho(\omega_1) - \rho(\omega_3)}{\rho(\omega_2) - \rho(\omega_3)} < 0.$$

To summarize, if $(x, 1)$ is a direction along which g is strictly convex, then

$$\begin{cases} x > 0 & \text{if } \frac{\rho(\omega_1)-\rho(\omega_3)}{\rho(\omega_2)-\rho(\omega_3)} < 0 \ (\iff \frac{s_{13}}{s_{23}} < 0), \\ x < 0 & \text{if } \frac{\rho(\omega_1)-\rho(\omega_3)}{\rho(\omega_2)-\rho(\omega_3)} > 0 \ (\iff \frac{s_{13}}{s_{23}} > 0). \end{cases} \tag{2.13}$$

[21] At least one of these terms is non-zero due to the assumption that $s_{12} = s_{23} = s_{13} = s \notin (0,2)$ does not hold.

By similar arguments, if $s_{13} = 0$,[22] the necessary condition for $(x, 1)$ being the direction along which g is strictly convex is

$$\begin{cases} x > 0 & \text{if } s_{23} > 0, \\ x < 0 & \text{if } s_{23} < 0 \end{cases}$$

and if $s_{23} = 0$, the necessary condition for $(x, 1)$ being the direction along which g is strictly convex is

$$\begin{cases} x > 0 & \text{if } s_{13} > 0, \\ x < 0 & \text{if } s_{13} < 0. \end{cases}$$

Given some interior prior, the sender splits it along a direction along which g is strictly convex and induces posteriors that lie on two edges of the simplex. We can distinguish the following cases:

1. If $\frac{s_{13}}{s_{23}} < 0$ or $s_{13} = 0 \wedge s_{23} > 0$ or $s_{23} = 0 \wedge s_{13} > 0$, then $x > 0$. Hence, the optimal split is either of the form $(q_1, 0, 1 - q_1)$, $(1 - q_2, q_2, 0)$ (pooling case (iii.a)) or of the form $(q_1, 1 - q_1, 0)$, $(0, q_2, 1 - q_2)$ (pooling case (iii.c)) depending on the prior.

2. If $\frac{s_{13}}{s_{23}} > 0$ or $s_{13} = 0 \wedge s_{23} < 0$ or $s_{23} = 0 \wedge s_{13} < 0$, then $x < 0$. In this case, we need to distinguish further:

   (a) If the optimal split goes along the direction $(-1, 1)$, it is of the form $(q_1, 0, 1 - q_1)$, $(0, q_2, 1 - q_2)$ (pooling case (iii.b)).

   (b) If the optimal split goes along direction $(x, 1)$ with $x < -1$, it is either of the form $(q_1, 0, 1-q_1)$, $(0, q_2, 1-q_2)$ (pooling case (iii.b)) or of the form $(q_1, 1-q_1, 0)$, $(0, q_2, 1 - q_2)$ (pooling case (iii.c)) depending on the prior.

   (c) If the optimal split goes along direction $(x, 1)$ with $x > -1$, it is either of the form $(q_1, 0, 1-q_1)$, $(0, q_2, 1-q_2)$ (pooling case (iii.b)) or of the form $(q_1, 0, 1-q_1)$, $(q_2, 1 - q_2, 0)$ (pooling case (iii.a)) depending on the prior.

Case (v). Proposition 2.1 applies and under Assumption 2.1 yields non-disclosure. □

[22] Notice that s_{13} and s_{23} cannot be simultaneously zero by assumption, because this would lead to case (v).

2.B Comment on Assumption 2.1

The structure of function g (see (2.5)) uncovered in Section 2.A.1 implies that for $n \geq 4$, there always exists a direction along which g is linear. Therefore, even when g is concave and non-disclosure is optimal, it is never uniquely optimal for $n \geq 4$. In particular, the sender is indifferent between sticking to the prior and splitting it to some posteriors from the space determined by the linear directions of g (and the prior), possibly all the way to the boundaries of the original simplex. Moreover, if g is concave, it is also concave on the boundary simplexes and we can repeat the same argument, proceeding downward in dimensions. For $n = 3$, by Proposition 2.1, g is concave only if it is linear in one direction. Hence, even for $n = 3$, non-disclosure is not uniquely optimal and the sender is indifferent between choosing a non-informative signal (keeping the belief at the prior) and splitting the prior into posteriors along the linear direction, all the way to the edges of the simplex. Therefore, pairwise signals (i.e., signals leading to posteriors supported on at most two states) are also always optimal.[23]

In the main text, we impose Assumption 2.1, which resolves indifference in favor of non-disclosure of states. It is a natural assumption that can be justified by the sender not wasting resources (time and energy) on communication when it is not needed (although the cost of communication is not featured explicitly in our model). This selection criterion simplifies the analysis. First, it enables us to avoid imposing some ad hoc assumptions about the selection of specific partial disclosure patterns from the indifference set. Second, a different natural assumption might be that the sender resolves her indifference in favor of splitting. However, this assumption would require us to impose some additional ad hoc assumptions about the selection of specific directions along which to split (for higher n) in order to deliver concrete predictions. Moreover, such a resolution of indifference would be very sensitive to the prior (even in terms of the predicted pooling structure), so we would need to keep track of the specific directions of indifference, which would render the analysis much more cumbersome.[24]

[23] This result is reminiscent of the result of Kolotilin and Wolitzky (2020) that there is no loss of generality from focusing on pairwise signals in their setup.

[24] To illustrate the dependence on the prior, for $n = 3$ under linear ρ (which is sufficient for global concavity or convexity), the direction of linearity is $(-\frac{\omega_3-\omega_2}{\omega_3-\omega_1}, 1)'$. Since the first component is strictly between 0 and -1, we can see that, while the non-disclosure is also optimal, the state-pooling structure (defined in Section 2.5) of the optimal informative signal can be either $\{\{\omega_1, \omega_3\}, \{\omega_2, \omega_3\}\}$ or

2.C Demonstration of the procedure for discovery of the state-pooling structure of the optimal signal

We demonstrate the application of the procedure for discovery of the state-pooling structure of the optimal signal (presented in Section 2.5) to the example introduced in Figure 2.4 (for convenience, we reproduce it in Figure 2.C.1 in this section). This demonstration is accompanied by Figure 2.C.2. *Red color* in Figure 2.C.2 represents highlighting as defined in Section 2.5 – final pools in full and candidate pools in dashed. *Green color* denotes cliques chosen for application of the non-disclosure test (Step 5 of the procedure).

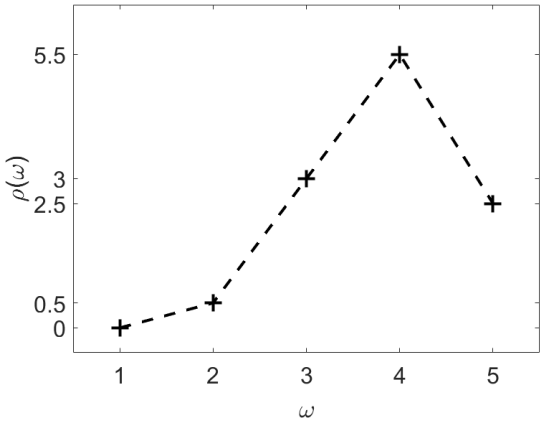


Figure 2.C.1: Preference misalignment function ρ considered for the demonstration of the graph procedure

The inputs to the procedure are the values of ω and $\rho(\omega)$ from Figure 2.C.1. From formula (2.4), we obtain the values of all s_{ij}: $s_{12} = 0.5$, $s_{13} = 1.5$, $s_{14} = \frac{5.5}{3}$, $s_{15} = \frac{2.5}{4}$, $s_{23} = 2.5$, $s_{24} = 2.5$, $s_{25} = \frac{2}{3}$, $s_{34} = 2.5$, $s_{35} = -\frac{1}{4}$, $s_{45} = -3$.

In (a) in Figure 2.C.2, we start with a fully connected graph on five nodes ($n = 5$) corresponding to states $1, 2, ..., 5$.

In (b) in Figure 2.C.2, we observe the same graph after the application of Steps 2 and 3 of the procedure. We removed all edges ij such that $s_{ij} \in (0, 2)$. As a result, node 1 became isolated, so we highlighted it in full. Hence, we can leave out node 1 from further analysis and focus on nodes 2, 3, 4, and 5.

$\{\{\omega_1, \omega_3\}, \{\omega_1, \omega_2\}\}$, depending on the prior.

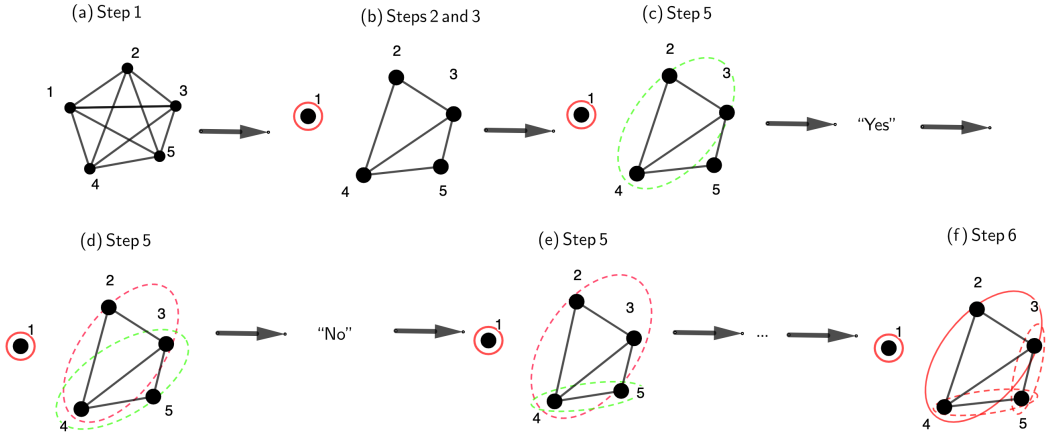


Figure 2.C.2: Illustration of the execution of the procedure, applied to the input from Figure 2.C.1; the output is in (f); red color represents highlighting as defined in Section 2.5 – final pools in full and candidate pools in dashed; green color denotes cliques chosen for application of the non-disclosure test

In (c) in Figure 2.C.2, we proceed to Steps 4 and 5 of the procedure. It is easily seen that there are two maximal cliques: one formed by nodes 2, 3, and 4 and one formed by nodes 3, 4, and 5. First, we inspect the maximal clique formed by 2, 3, and 4 (highlighted in green) and we apply the non-disclosure test. The non-disclosure condition holds, so we highlight the maximal clique $\{2, 3, 4\}$ in dashed (as illustrated in (d)). Hence, we do not need to consider any more of its subsets in Step 5 and we can move our focus to the other maximal clique.

In (d) in Figure 2.C.2, we inspect the maximal clique formed by nodes 3, 4, and 5 (highlighted in green). The non-disclosure condition does not hold, so we denote the maximal clique $\{3, 4, 5\}$ as inspected and proceed to consider its subsets of cardinality 2.

In (e) in Figure 2.C.2, we first consider the clique formed by nodes 4 and 5. As the non-disclosure condition is satisfied, we highlight this clique in dashed. Proceeding with the iteration, we test clique $\{3, 5\}$. Again, the non-disclosure condition is satisfied, so we highlight it in dashed. Finally, clique $\{3, 4\}$ is a subset of the highlighted set $\{2, 3, 4\}$, so we do not test it.

In (f) in Figure 2.C.2, we proceed to Step 6 of the procedure: as node 2 belongs to only one highlighted clique, $\{2, 3, 4\}$, we highlight that clique in full. The output of the procedure is

depicted in (f) in Figure 2.C.2: the singleton pool $\{1\}$ and pool $\{2,3,4\}$ highlighted in full and pools $\{3,5\}$ and $\{4,5\}$ highlighted in dashed. Hence, the posteriors induced by the optimal signal certainly include a posterior supported on states $\omega_2 = 2, \omega_3 = 3, \omega_4 = 4$ and the posterior δ_{ω_1}. Moreover, the optimal signal will induce at least one posterior supported on $\omega_3 = 3, \omega_5 = 5$ or $\omega_4 = 4, \omega_5 = 5$.

Chapter 3

Discrimination in Disclosing Information about Female Workers: Experimental Evidence

Abstract

Sona Badalyan, Darya Korlyakova, and Rastislav Rehák[1]

We focus on communication among hiring team members and document the existence of discrimination in the disclosure of information about candidates. In particular, we conduct an online experiment with a nationally representative sample of Czech individuals who act as human resource assistants and hiring managers in our online labor market. The main novel feature of our experiment is the monitoring of information flow between

[1]This chapter is joint work based on Badalyan, S., Korlyakova, D., and Rehák, R. (2023) "Disclosure Discrimination: An Experiment Focusing on Communication in the Hiring Process," CERGE-EI Working Paper Series No. 743. *Author contributions*: Badalyan, S., Korlyakova, D., and Rehák, R. designed the experiment, collected and analyzed the data, and wrote the paper. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101002898). This study was supported by Charles University, GAUK project No. 333221. We thank Michal Bauer, Julie Chytilová, Filip Matějka, Andreas Menzel, Nikolas Mittag, and Jaroslav Groero for valuable discussions. We are also grateful to Data Collect and MEDIAN for excellent cooperation on data collection. The study was pre-registered in the AEA RCT Registry: AEARCTR-0008662. The experimental design was approved by the CERGE Ethical Committee.

human resource assistants and hiring managers. We exogenously manipulate candidates' names to explore the causal effects of their gender on information that assistants select for managers. Our findings reveal that assistants disclose more information about family and less information about work for female candidates than for male candidates. An in-depth analysis of types of information disclosed suggests that gender stereotypes play an important role in this disclosure discrimination.

Keywords: Information, Disclosure, Hiring, Discrimination, Online Experiment
JEL Codes: C90, D83, J71

3.1 Introduction

Information about job applicants is a key input that firms use when making hiring decisions. It has long been acknowledged that the lack of access to individual-level information can lead to statistical discrimination against certain societal groups (Phelps, 1972). More recently, researchers have become interested in understanding mechanisms that may underlie biases in information acquired depending on the group characteristics of job applicants, which could arise even when individual-level information is available. In particular, Bartoš et al. (2016) show that employers may discriminate in attention allocation in the presence of cognitive constraints.

In this paper, we focus on *disclosure discrimination*—biases that arise due to the exchange of information among individuals in hierarchical organizations.[2] For example, in communication with a hiring manager, human resource (HR) assistants may emphasize strong features of a majority applicant and make them less salient in the case of a minority applicant. Assistants could also omit some information about some applicants to promote a candidate whom they favor.

Our primary question is whether HR assistants select different information for hiring managers depending on the applicants' gender. One reason this question is understudied

[2]The importance of smooth communication between human resource specialists and hiring managers is a popular topic on many career-related websites. For instance, Glassdoor (2021) advises that "Recruiters and hiring managers must be in constant contact with one another to be effective and efficient - and a great way to do this is to hold post-interview debriefs via phone, Skype, or in person. [...] if feedback is delayed or non-existent, [...] hiring decisions [...] can be postponed."

in previous academic work is that monitoring communication during a hiring process is difficult, especially in field settings. Nevertheless, this topic requires attention because recent evidence is indicative of possible discrimination in information transmission in the hiring context. Specifically, Kline, Rose, and Walters (2022) find that firms with greater centralization of recruiting—a measure indicative of hiring responsibility being divided among fewer individuals—have smaller racial and gender callback gaps. Moreover, a meta-analysis by Quillian, Lee, and Oliver (2020) shows that discrimination at the interview stage contributes substantially to less-frequent job offers to racial minorities than to majority candidates. Although interviewers are not necessarily responsible for final hiring decisions, they may affect them by sharing and emphasizing observations with hiring managers, which is plausible according to our qualitative interviews with HR specialists. For example, Rivera (in Dobbin and Kalev, 2016) finds that unsuccessful test results of female and African American candidates are scrutinized more relative to those of White men during hiring meetings. Assistants may also want to accommodate the biased preferences of the hiring team and thus manipulate the disclosure of their information about potential employees accordingly.

To address our research question, we conduct an online experiment with a large nationally-representative sample of Czech individuals (N=757) who act as HR assistants.[3] These individuals select information from eight workers' profiles, which contain details about the workers' demographics, education, professional experience, qualifications, and personal qualities. To exogenously manipulate gender, we randomly assign names to the profiles. The random assignment of names also aims to vary the workers' nationality, which we will discuss in an extended version of our paper. To gain additional insights into the mechanisms that may lead to potential discrimination in disclosure, we collect data on assistants' attention during the information-selection task. While the assistants only select information about the workers, we recruit a different sample of participants in the experiment to act as hiring managers, who make final hiring decisions (specifically, the workers can be hired for a financial task). Importantly, before making each hiring decision, a manager sees information that an assistant has disclosed about each worker in

[3]Women are likely to be over-represented among HR assistants. According to the International Labor Organization (2020), 61% of other clerical support workers, a category that includes HR assistants, are women. Nevertheless, we decided to over-sample men (relative to their potential share in the HR profession) to have higher power to detect gender differences.

addition to the manipulated name. Eventually, the managers can reward the assistants for selected information if they find the selection valuable. The managers want to learn the potential of workers on the financial task, because their performance on this task affects the managers' payoffs.

Cleanly identifying the causal effects of gender on disclosure is empirically challenging if the content of candidates' profiles differs depending on gender. For this reason, we showed different assistants the same profiles with exogenously varied names. These profiles are real; we constructed them on the basis of information collected in a pre-experimental survey. We refer to the participants of this survey as workers because they performed real-effort tasks. We had to assign the names to the profiles exogenously because it was practically impossible to match precisely real information-rich profiles of men and women that would feature their actual names. The exogenous assignment of names is commonly used by correspondence studies (see, for instance, Bertrand and Duflo [2017] for a review) in which researchers send the same fictitious applications with different names to real firms. Employers in these studies are not informed that the applicants are fictitious so that they behave in a realistic manner and have an incentive to study the information about those applicants. Similarly, we omit the information that the workers' names are fictitious to make the assistants take the information-selection task seriously.

Our choice of an online experiment as a suitable method to explore patterns in assistants' disclosure is inspired by recent experimental literature. Identifying the components of discrimination or mechanisms that may contribute to group-based disparities serves as a basis for successful policy responses, but is often unfeasible in natural settings. In this regard, researchers started to conduct hiring experiments on online crowdsourcing platforms in which they assign participants the roles of workers, recruiters, hiring managers, or employers (Bohren, Hull, and Imas, 2022; Bohren et al., 2019). Furthermore, a growing number of studies (Cappelen, Falch, and Tungodden 2019; Almås, Cappelen, and Tungodden 2020; Cappelen et al. 2020) address questions related to distributional preferences by observing the decisions of impartial spectators with respect to workers' outcomes in online settings. These studies often recruit participants with the help of research agencies similar to those we cooperated with.

Our main findings are the following. First, if a CV has a female name, assistants select more demographic information for hiring managers, in particular information that may signal increased household responsibilities. For example, assistants are 31.4% more likely

to disclose information about the number of children for female workers than for their male counterparts. The effects are driven by male assistants and are somewhat stronger for those who seem to be more supportive of traditional gender roles. Second, assistants provide less work-related information about female workers. This effect is driven by our low-quality profiles. However, this overall negative effect hides important differences across types of information: whereas assistants disclose less information about the job responsibilities of female workers with low-quality profiles, they provide more information about their job positions.

The differential disclosure of information depending on candidates' gender seems to be connected to gender stereotypes. By providing more information on the number of children and marital status of women compared to men, the assistants emphasize (consciously or unconsciously) the importance of family for women. This information could make family obligations salient, which can reduce the chances of women finding a job (Becker, Fernandes, and Weichselbaumer, 2019; Petit, 2007). By over-providing information about job positions in low-quality profiles for female workers relative to male workers, assistants highlight women's stereotypical occupational choices, because our low-quality profiles tend to represent workers from female-dominated professional fields.

A distinctive contribution of our online experiment is the heterogeneity analysis on the characteristics of assistants. On the methodological side, such analysis is difficult to perform in standard correspondence studies (e.g., Quillian et al., 2017; Kaas and Manger, 2012; He, Li, and Han, forthcoming; Bertrand and Mullainathan, 2004) because to collect information about employers' demographics, researchers would need to ask them additional questions, which would make the employers aware of their participation in an experiment. On the theoretical side, heterogeneity analysis reveals systematic differences in disclosure patterns across assistants with different characteristics, e.g., gender. This suggests that the composition of a hiring team affects communication. Therefore, the role of an HR assistant cannot be formally reduced to acting merely as an attention system of a manager, which could be captured by a single-agent model.[4]

In addition to correspondence studies, our paper adds to other types of experiments on

[4]If there were no systematic differences in disclosure across assistants, the situation could be modeled parsimoniously as if it was directly the hiring managers directing their attention to the disclosed pieces of information (or just asking the assistants to prepare those pieces of information without the assistants' subjective involvement in the selection process). We thank Filip Matějka for this observation.

discrimination in hiring. In vignette studies (e.g., Kübler, Schmid, and Stüber, 2018; Bertogg et al., 2020; Oesch, 2020), professionals (often human resource managers) evaluate fictitious candidates' CVs in terms of the likelihood that they would invite the candidates to an interview—the next stage of the recruitment process—or consider them for a specific job. While these experiments rely on subjects' hypothetical choices, Kessler, Low, and Sullivan (2019) design an incentivized resume rating. In their study, employers express interest in hiring hypothetical candidates, knowing that these choices reveal their preferences which will be used to match them (the employers) with actual candidates. Our work is different from these types of experiments because we incorporate the involvement of multiple decision-makers in hiring in order to reflect more closely the real-life processes. Communication among the decision-makers could be a channel through which discrimination propagates and unfavorable stereotypes emerge.

Our results about the role of workers' gender are broadly related to recent evidence suggesting that gender discrimination often manifests itself in subtle forms (e.g., Dupas et al., 2021; Hengel, 2022). Many of these studies (Barron et al., 2022; Brock and De Haas, forthcoming) aim to detect implicit gender bias that does not materialize in simple decisions. For instance, Brock and De Haas (forthcoming) do not observe that loan officers discriminate against women directly: unconditional loan approval rates are the same for male and female applicants. However, female applicants are 30% more likely to be asked for a guarantor. We contribute to this literature by identifying a subtle form of disclosure discrimination. Our assistants do not seem to systematically provide unfavorable information about one group of workers (e.g., low education, a limited set of skills relevant for the hiring task, below-average performance on previous real-effort tasks, or self-reported weaknesses). However, they tend to emphasize women's family situations and their employment in traditionally female occupations.

Our uncovered gender discrimination in disclosure also relates to the literature on the role of stereotypes in governing the decision-making of employers, recruiters, and other professionals (e.g., Wu, 2018; Gallen and Wasserman, 2021). Within this literature, a few studies (González, Cortina, and Rodríguez, 2019; Van Borm and Baert, 2022) find that gender stereotypes are triggered more strongly when female CVs explicitly mention family responsibilities and that gender bias in recruitment becomes stronger if female candidates have children. We contribute to this literature by uncovering a new domain in which gender stereotypes may influence decisions: selection of candidates' information

by HR specialists for later stages of the recruitment process.

The over-provision of family-related information and information about female-dominated positions for women is reminiscent of the representative signal distortion (RSD) documented by Esponda, Oprea, and Yuksel (2023). An evaluator of a CV using RSD looks for evidence representative of the group the candidate belongs to. For example, if women are more likely to be cashiers, the evaluator takes such a piece of information into account as more likely for women than for men, which biases subsequent inference. However, the perception of the evaluator that guides RSD is unobserved. In contrast, we observe explicitly information selection with patterns that can be explained by representativeness: women are more likely to be employed in female-dominated jobs (by definition) and they are more likely to be responsible for taking care of a household and children. This suggests that representativeness might be driving not only inference in decision making (as in Esponda, Oprea, and Yuksel [2023]), but also communication.

In the closest paper to ours, Eberhardt, Facchini, and Rueda (2022) investigate which attributes recommendation letter writers emphasize when describing academic job-market candidates of different genders. The authors find that women are more frequently described using "grindstone" terms (e.g., "hard-working" or "dedicated") while also less likely praised for their ability. Our findings also suggest that individuals aim to emphasize somewhat different characteristics of female job seekers by means of differential disclosure. Our paper complements Eberhardt, Facchini, and Rueda (2022) in several ways. First, we provide clean identification in a stylized experimental setting, while they use machine-learning techniques to uncover tendencies in real-world data. Specifically, we study causal effects of candidates' gender on information selection, while Eberhardt, Facchini, and Rueda (2022) measure associations between job-market candidates' gender and language used in their reference letters. Second, the agents who choose information in our setting are HR assistants representing the labor-demand side of the market, while the supervisors writing the reference letters in Eberhardt, Facchini, and Rueda (2022) represent the labor-supply side.

The rest of this chapter is organized as follows. First, we describe the study design and our samples. Second, we present our identification strategy. Finally, we discuss our experimental results and conclude.

3.2 Study design

In this section, we describe the online experiment with a representative sample of Czech respondents, to whom we assign the role of HR assistants in order to test for discrimination in information disclosure. We also outline two supplementary surveys that were conducted (i) to collect information for workers' profiles and (ii) to provide assistants with real incentives.

Figure 3.A.1 in the Appendix provides an overview of the project and Figure 3.A.2 focuses on the flow of the main experiment with assistants.

3.2.1 Sample of assistants

We hired subjects for the assistant role with the help of Data Collect, a local research agency, by using their online panel. The data were collected from a sample of 757 adults during November-December 2021. The sample is representative of the Czech general population aged 18-64 years in terms of gender, age, education, and regional coverage (Table 3.B.1). The characteristics of the assistants are summarized in Table 3.B.2. About 13% of the assistants report having recruitment experience.[5]

After the main part of the experiment (the information selection task described below), we asked the assistants how much they had thought about a hiring manager during the information-selection task. The answers were coded on an 11-point scale, where 0 means "not at all" and 10 means "a lot." The average score is 8.15 (83% of assistants chose 7-10), which suggests that the manager's role in the information-selection process of our experimental subjects is high.

A number of additional measures suggest that the assistants largely took the task seriously. A median assistant spent about 11.5 minutes on selecting information from the 8 profiles. The assistants tended to disclose more than half of a worker's profile and to provide diverse information about a worker.

[5]We report the main regression results in the subsample of assistants with recruitment experience in Tables 3.B.14, 3.B.15, and 3.B.16. They should be compared with Figure 3.1 and Tables 3.1 and 3.2, respectively, which are based on the whole sample of assistants.

Before providing the details about the assistants' main task, we explain how the workers' profiles, from which the assistants selected information, were constructed and which elements they included.

3.2.2 Creating workers' profiles

To collect information for workers' profiles, we conducted a survey with 20 Czech respondents with the help of MEDIAN, a different research agency. This survey consisted of real-effort tasks and questions about demographics, education, work experience, etc. To reduce workers' fatigue, we asked MEDIAN for additional information (e.g., media consumption and self-reported financial literacy) on the same respondents from the agency's previous surveys. Before asking for consent to participate in our survey, we explicitly informed respondents that we may use their data when creating questionnaires for other respondents but these data would never be linked to their names or other identifying information.

We aimed to create a diverse set of credible profiles, which would resemble real-life CVs or LinkedIn profiles (we describe the content of the profiles below). In particular, we had to ensure that the profiles did not contain suspicious information, especially when varying the names attached to them—for example, we did not want to use a profile of a construction worker because we could not credibly assign a female name to it. The goal was to make the task for the assistants realistic and engaging. In the end, we chose 8 workers whose responses and task results were used to construct the 8 profiles.

The 8 workers were being hired for an actual task with a series of financial decisions (we describe the hiring managers' task in a separate section later). The assistants were aware of this and the 8 profiles were constructed to be quite informative about the workers' qualifications for this task. The financial task consisted of 10 multiple-choice questions which involved both computational skills and financial knowledge. For example, the workers were asked to calculate the balance on a savings account after a year given the initial balance and the interest rate. In another question, they had to indicate the most volatile asset in a given list.

The content of all profiles is in Appendix 3.C (page 132 onwards). Here, we describe the sections featured in the profiles:

Summary. This section describes the workers' self-reported personal strengths, weaknesses, and their opinion about their own financial skills or skills that they find important (e.g., "learning new things").

Demographics. This section includes mostly information about the workers' demographics—age, marital status, and number of children. It also provides information about whether the worker has a driving license and how many surveys he or she has completed in the past (based on the agency's records).

Education. This section provides information about the workers' level of education, field of studies, and favorite subjects (e.g., Math, Risk Management).

Work. This section provides information about the workers' job sector, current position, years of experience in the current role, and job responsibilities (e.g., communication with governmental offices, database administration). In the case of one profile, we refer to the previous position instead of the current one because the worker is not employed. We truthfully mention that this worker is on parental leave.

Certificates. This section summarizes the workers' results on three real-effort tasks that should signal their abilities in mathematics and finance and general effort. In the mathematical task, workers were asked to answer 10 mathematics questions within a limited time. The questions are inspired by those of Bohren et al. (2019), for example: (i) "Which of the following is an integer multiple of 11?" (ii) "$16 < x+8 < 26$. Which of the following could x be?" The workers always chose from four options. In the financial knowledge quiz, the workers were asked to answer 5 multiple-choice questions that aimed to test whether they understand the concepts of inflation, exchange rate, company shares, etc. When preparing this task, we adapted examples from the Czech National Bank and other sources with financial literacy tests. In the slider task, which is frequently used in the experimental literature (e.g., Gill and Prowse, 2019; Bradler, Neckermann, and Warnke, 2019; Gill and Prowse, 2012), the workers had to position 48 sliders at the exact position of 50 during a limited time. Each slider was initially positioned at a random number between 0 and 100.

We chose these tasks because we hypothesized that the assistants would disclose information depending on its relevance for the hiring task. A priori, the financial knowledge quiz seemed to have the highest predictive power for the workers' performance on the task

with a series of financial decisions, while the slider task seemed to be the least relevant.

Judging the workers' performance on the three tasks without a reference point would be difficult for the assistants, especially in the case of the first profile that assistants would see. Thus, we included the average score of all workers who took part in the survey for each task.

Volunteering.[6] This section informs about the workers' observed donations for a good cause. MEDIAN provided us with the data on the frequency of workers' donations in past surveys. Each time their respondents completed a survey, they were redirected to the agency's page where they had to decide whether their survey completion fee should be transferred to their bank account, donated to a charity from a list, or whether they wanted to give it up. If a worker chose to send his or her fee to a charity in the past, we mention on his or her profile in what percentage of surveys the worker made the decision to donate. Furthermore, at the end of our survey with the workers, we asked the participants whether they would like to complete another survey in the upcoming days and donate a fee from participating in that survey to a charity of their choice. If a worker chose "yes" and MEDIAN later confirmed that the worker chose to donate his or her money *after* filling in the other questionnaire, we mentioned the worker's donation decision in his or her profile.

Skills. This section enumerates the workers' self-reported skills, such as Microsoft Office experience, English language proficiency, familiarity with online banking, experience with data analysis, customer service, product management, and so on. We included the information about online banking because we expected that the assistants may relate it to financial literacy and thus to the workers' performance on the hiring task.

Interests. This section provides information about the workers' leisure-time activities and interests, e.g., sports, traveling, or reading news about finance/business/economics in newspapers or on the Internet.

We populated each section of a profile only with true information gathered from the same worker. Since our workers could decide how many details to provide about themselves (in our survey and previous surveys with the data collection agency), the resulting 8

[6]This section is missing in 4 profiles because we found it hardly realistic that individuals would voluntarily report that they never donated to a charity.

profiles differ somewhat in length. Specifically, they contain between 24 and 35 pieces of information.

Surveying workers with diverse educational and professional backgrounds enabled us to construct "low- and high-quality" profiles. We associate profile quality with the worker's suitability for the financial (hiring) task. As previous research has documented a positive correlation between a person's financial literacy and education (Lusardi, Mitchell, and Curto, 2010), we categorize profiles as low-quality if they come from the workers who completed at most secondary education, while the high-quality profiles come from the workers with a university degree.[7] Half of the profiles are classified as low-quality.

The low- and high-quality profiles differ along several other dimensions besides education. In particular, the low-quality profiles represent mostly workers from low-skilled occupations, whose self-reported skills and job responsibilities tend to signal that they are less-suitable candidates for the financial task.[8] Moreover, the low-quality workers do not use online banking, report only partial knowledge of English (compared to good knowledge for the high-quality ones), and made no charity donations. An example of a high-quality profile is Ondřej's profile in Appendix 3.C; an example of a low-quality profile is Lucie's profile in Appendix 3.C.

[7]Heterogeneity along the quality dimension is an important element of our experimental design because we might expect differential treatment of female workers with lower qualifications. For instance, Bohren, Imas, and Rosenberg (2019) ran an experiment on a large online platform in which they observed strong discrimination against female users with novice accounts but favorable treatment for women with a history of positive reviews.

[8]Note that both types of profiles include "positive" as well as "negative" information. This is natural given that we used real data. However, the low-quality profiles contain more information that may put a candidate at a disadvantage compared to the high-quality counterparts. An added value of having profiles with "mixed" information is that such ambiguity might reveal implicit discrimination (Cunningham and de Quidt, 2022). For example, due to self-image or social-image concerns, an assistant may be reluctant to select solely unfavorable information about a worker whose group the assistant dislikes or finds less competent. However, disclosing the worker's weaknesses together with less relevant positive characteristics could help the assistant disguise his or her bias.

3.2.3 Experiment with assistants

We remind the reader that Figure 3.A.2 in the Appendix provides a depiction of the flow of the experiment with the assistants. The full instructions for the assistants (translated from Czech) can be found in Appendix 3.C.

Instructions, incentives, and the information selection task

In the beginning, the subjects were informed that they would act as assistants for recruiting workers in our online labor market. We emphasized that this is not a traditional survey that asks about hypothetical situations and that their decisions may have real financial consequences for other respondents.

Next, the assistants learnt that they would see 8 CVs and their task would be to select information they would like to disclose to another survey participant, who would act as a hiring manager. The assistants knew that the hiring manager would see only the information disclosed about a worker, along with the name on the CV, when making the hiring decision for the financial task. If an assistant decided not to disclose any information about a worker, the manager would see only an empty profile with a name.

We incentivized the assistants to take the disclosure task seriously in the following manner. If a manager found the disclosed information useful, he or she could allocate to the assistant an additional bonus of up to 50 Czech crowns ($\sim$ \$2); this bonus did not cost the managers anything (it was a pure reward) and the assistants knew that. Furthermore, the assistants knew that the managers would make multiple hiring decisions during a limited time, so the simplified versions of the CVs would be of great help to them. Finally, the assistants were informed that the managers would benefit financially from hiring workers with good performance on the financial task. Hired workers would also earn additional money.

We included a comprehension check at the end of the instructions. Specifically, we aimed to test the assistants' general understanding of (i) their task, (ii) the managers' role and the information available to them, and (iii) the incentives that they (assistants) have. The assistants had to evaluate whether each of three statements was true or false in order to proceed to the information selection task. We showed the correct answers on the next page along with a scheme summarizing the key points of the instructions.

In the main task, each assistant selected information from the same set of 8 different profiles, which were shown sequentially and their order was randomized. To indicate the selection, the assistants had to tick information they wanted to send to a manager directly in the CVs. As a default, no specific information was preselected, i.e., the assistants had to actively select what to disclose. There was no limit on the amount of information pieces the assistants could select. After the assistants selected information from each profile, we showed them a preview of what a manager would see about a specific worker based on their selection. We allowed the assistants to return to the previous page to change their disclosure choices.

Treatments

To study the effect of workers' gender on assistants' disclosure, we randomly assigned a name to a profile to form a CV (independently across profiles and assistants).[9] Orthogonally to gender, we varied workers' nationality. Hence, we used a 2x2 design; an assistant could potentially see a profile in four different versions: local male, local female, foreign male, and foreign female. To mitigate the effect of specific names, each profile had a different set of names that could be attached to it. The full list of names is presented in Table 3.B.3.[10,11]

To summarize and pin down our nomenclature, a *profile* is a nameless set of information representing a real worker and a *CV* is a profile with a fictitious name attached to it. Each assistant sees the same 8 profiles (in a random order).

The assistants were not informed that workers' names were fictitious. Including this information could make the subjects suspicious about the real gender (and nationality)

[9] To cleanly identify the causal effects of the workers' gender on disclosure, we had to compare CVs with different names but the same information content. However, it was practically impossible to construct identical profiles based on data from different workers because CVs contained numerous pieces of information. For this reason, we assigned fictitious names to real profiles.

[10] We also displayed the IDs of workers, invented by us, next to the workers' names to substitute for the lack of surnames and to make the task more realistic. Our IDs do not reveal the identity of the real workers.

[11] Our further discussion and analysis are more narrowly focused on the importance of workers' gender for the assistants' disclosure decisions. The nationality dimension, whose effects are still being explored, will be presented in the extended version of our paper.

of the workers behind the profiles, which would introduce a confound difficult to control for. Moreover, it could jeopardize our effort to make the main task as realistic and important as possible and reduce the assistants' effort.

Our manipulation of attributes of interest with the help of a first name is somewhat less salient compared to previous literature on discrimination, which uses both a first name and a surname. We did not use the surnames because we were concerned that the assistants may think that we disrespected the workers' anonymity by providing personally identifiable information.

We included a manipulation check to test whether our treatment was salient enough. Specifically, after the assistants finished information selection from the last CV, we asked them about the gender of that last worker. At this stage, the assistants could not return to the last CV to check the name. We did not inform the assistants beforehand that we planned to check their attention later to avoid the experimenter demand effect. For the same reason, we did not include a manipulation check after each CV; only the last one. Correct answers on the manipulation check were incentivized by an extra bonus. We observe that 92% of assistants accurately identified the gender of their last CV.

Tables 3.B.4 and 3.B.5 demonstrate that the randomization was successful, i.e., the treatment arms are well balanced and the observables are jointly unrelated to the treatment status.

Outcomes

Capturing communication in a disciplined manner is difficult. Even the simple form of communication that we restrict to—disclosure—results in a large amount of possible patterns. To avoid data mining, we pre-specified to inspect a small set of outcome variables: https://www.socialscienceregistry.org/trials/8662.

Disclosure-related outcomes. We adopt a "top-down" approach to study the effects of workers' gender on disclosure. This means that our primary outcome of interest is the overall share of information pieces that an assistant discloses from a CV. Subsequently, we study the shares of disclosed information pieces in the sections described above (e.g., Demographics, Education, Work). If the treatment significantly affects disclosure from a specific section, we take a closer look at the content of this section to understand which

pieces of information drive the effect. For example, if we observe a treatment effect on disclosure from the Demographics section, we additionally compare how frequently assistants disclose information about workers' age, marital status, number of children, driving license, and number of completed surveys for men vs. women.

Attention-related outcomes. To study possible drivers of (potential) disclosure discrimination, we additionally collected data on assistants' attention allocated to CVs. Specifically, we recorded the time that each assistant spent on selecting information from each CV. As we did not impose any limit on time that assistants should spend per CV, the subjects could move through CVs as quickly as they wanted. We also measured how frequently assistants chose to learn more about some specific pieces of information in the CVs. For this purpose, we embedded 4-6 buttons in each profile (in sections Demographics, Education, Work, Certificates, and Volunteering), next to information pieces that may not be self-explanatory, potentially causing assistants to be interested in further details. For instance, a button next to the slider-task results (section Certificates) informed assistants about the nature of this task if the person clicked on it: *The slider task is a mechanical task in which participants had to center within a 2-minute limit as many sliders as possible (max. 48) with a random initial position.* The content and position of these "learn-more" (or "more information") buttons within the profiles can be seen in Appendix 3.C (page 132 onwards). Our outcome variable is the total number of assistants' clicks on the "learn-more" buttons in a CV. This measure also captures repeated clicks on the same button.

3.2.4 Managers' hiring decisions

After running the experiment with the assistants, we conducted a large-scale survey with a different sample of respondents who acted as hiring managers. The purpose of this data collection, which was performed in cooperation with the same research agency (Data Collect), was twofold. First, it was necessary to conduct this survey not to deceive our experimental subjects. We promised the assistants that information that they would select about the workers would be shown to another survey respondent and that this respondent would decide how to reward their effort. Second, we intended to get insight into the consequences of potential discrimination in disclosure. In this dissertation chapter, we focus only on the experiment with assistants and we leave the discussion of the results

from the managers' survey to the extended version of our paper.

Each manager was matched with a random assistant[12] and saw information that the assistant selected from the 8 CVs. The order of CVs was re-randomized. For similar reasons as for the assistants, the managers did not know that the names were fictitious. Immediately after a manager saw a CV pre-processed by an assistant, she/he made a hiring decision about the corresponding worker.[13] After a manager made all 8 hiring decisions, she/he could reward the assistant whom they were paired with by a real bonus if they found the assistant's selection of information useful.

3.3 Identification

To quantify the effect of gender on disclosure of information passed by an assistant to a manager, we employ the 2-way Fixed Effects Model. Each assistant i sees 8 profiles indexed by j.

Baseline regressions

We start by estimating the following regression model:

$$Y_{ij} = \eta + \tau T_{ij}^{FEM} + \mu_i + \phi_j + \xi_{ij}. \tag{3.1}$$

Y_{ij} is an outcome variable (e.g., share of disclosed information pieces by assistant i in profile j). T_{ij}^{FEM} is an indicator of whether assistant i saw profile j with a female name. We control for unobservables fixed over assistants and profiles by including assistant fixed effects μ_i, as well as dummies for the profiles ϕ_j. The coefficient of interest is τ; it shows the effect of female name on assistants' disclosure or attention.

Heterogeneity

We are also interested in whether the treatment effects differ for subgroups of assistants

[12]By chance, a few assistants were paired with two managers. In these cases, we randomly chose one of the assigned managers and recorded his or her decision while calculating the extra reward to the corresponding assistant. Consequently, we had to recruit additional managers to reward the unmatched assistants.

[13]Hiring decisions were incentivized with a small bonus, which was increasing in the worker's performance on the task with a series of financial decisions.

with different characteristics (in particular, assistants' with different gender or attitudes toward women) and for profiles of different quality. To examine these heterogeneous effects, we augment equation (3.1) by including an interaction of the treatment indicator with a heterogeneity variable of interest.

Clustering

In all models, we cluster the errors at the assistants' level to address potential correlation across profiles.

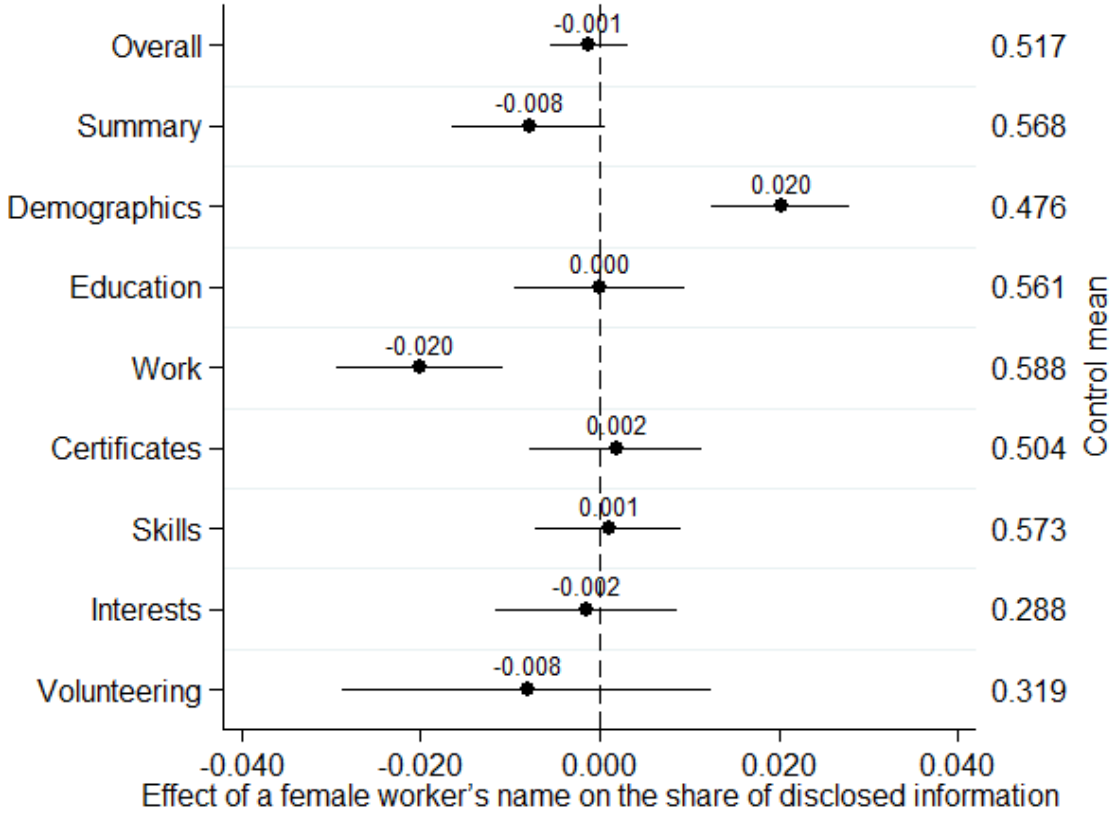
3.4 Results

This section presents the results from our experiment with the assistants. Specifically, we discuss that assistants seem to rely on gender stereotypes when disclosing information about female workers.

Figure 3.1 illustrates the causal effects of a female name on the share of disclosed information from a worker's whole CV and from the particular sections. The sizes of the control means indicate that the assistants tend to provide a nontrivial amount of information about the workers and their disclosure covers a diverse set of profile sections. The assistants select on average 51.7% of information (around 16 pieces) from a male CV. The assistants disclose the most about the male workers' work experience, self-reported skills, personal qualities, and education while they tend to neglect the information about the workers' interests and volunteering activities. Assigning a female name to a profile significantly increases the amount of information disclosed from Demographics and decreases the amount of information disclosed from Work. In particular, the assistants disclose on average 2 percentage points (pp) more information from Demographics for a female CV than for a male CV ($p<0.01$; 4.2% increase relative to the control mean[14]). At the same time, they select on average 2pp less information from Work for a female CV than for a male CV ($p<0.01$; 3.4% decrease). To gain deeper insights into these treatment effects,

[14]When presenting percent changes throughout this paper, we always compare the treatment effects to the control mean (i.e., the average value of the outcome in the male group) or to the control mean in a specific subsample (e.g., male assistants) in the case of heterogeneity analyses. We omit the description of the baseline group in the text, but its specification can be found in the notes for the corresponding tables or figures.

Figure 3.1: Effect of a female name on the disclosure of information (overall and from each section)



Notes: Coefficient plots. Each row corresponds to the regression of the share of disclosed information in the corresponding category (left axis) on the indicator of female name on a CV (with assistants' and profiles' fixed effects). The points represent the estimated coefficients and the bars represent the 95% confidence intervals. The control means (right column) are simple means of the share of disclosed information in the corresponding category over CVs with male names.

we study which pieces of information drive these differences and run the pre-specified heterogeneity analyses for each of the two profile sections.

3.4.1 Workers' gender and disclosure of demographic information

Table 3.1 shows the results of regressions in which all information pieces from Demographics serve as dependent variables. Assistants are 2.4pp more likely to disclose information about marital status and 8.2pp more likely to disclose information about the number of children if a CV has a female name ($p<0.01$ in both cases). This corresponds to an increase of 6.3% and 31.4%, respectively, compared to the control means. The

finding that assistants provide family-related information more frequently in the case of female workers[15] suggests that they may find it more relevant for hiring women. Correspondence studies (e.g., Becker, Fernandes, and Weichselbaumer, 2019; Petit, 2007) systematically document that hiring discrimination against women prevails among those applicants whose demographics signal a higher likelihood of becoming pregnant or overoccupied with childcare. Hence, the tendency to signal this kind of information for women (even in our online context) suggests its prominent role in discrimination against women.

Table 3.1: Effect of a female worker's name on the disclosure of Demographic information

	(1) Age	(2) Marital status	(3) Children	(4) Driving license	(5) Surveys
Female	-0.005	0.024***	0.082***	-0.003	0.000
	(0.006)	(0.008)	(0.010)	(0.006)	(0.007)
Control mean	0.753	0.384	0.261	0.715	0.250
Observations	6056	5299	6056	6056	6056

Note: Regressions of specific demographic information pieces on the Female treatment indicator. *Surveys* contains the actual number of surveys that a worker completed in the past. All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. The control means are the average values of the outcomes in the male-CVs group.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Next, we discuss whether the Female treatment effects on the disclosure of workers' demographics differ among different subgroups of assistants or workers' profiles. In one of these analyses, we split the sample by assistants' bias against women. We constructed this variable in the following manner. At the end of our experiment, we asked the assistants to what extent they agree or disagree with different statements in relation to gender roles and stereotypes. For instance, they had to express their (dis)agreement with whether women should be more responsible for household chores than men or whether boys are more talented in technical subjects and mathematics relative to girls. The assistants answered on a 5-point Likert scale where 1 stood for "fully agree" and 5 represented

[15] In one of the profiles, we (truthfully) mention that the worker is on parental leave. The assistants are 1.6pp more likely to disclose this information for a female CV (p=0.65; 2.6% increase compared to the control mean). The results are presented in column 1 of Table 3.2 and are based on the OLS regression (N=757) in which the assistants' characteristics (age, gender, household size, educational and regional dummies, and a recruitment-experience dummy) are included.

"fully disagree." To construct an index indicating tolerance to women, we first ensured that higher values always encode "better" perception of women and then calculated the average of each assistant's responses to all statements. For ease of interpretation in the heterogeneity analysis, we use a dummy variable (called "biased against women") equal to one if the value of the tolerance index is less than or equal to the median.[16]

Only men reveal significantly more demographic information if a CV has a female name, as demonstrated in column 1 of Table 3.B.6. A male assistant selects 3.2pp more demographic information about a female worker relative to his average disclosure of 49.4% from Demographics in the case of a male worker ($p<0.01$; 6.5%). In contrast, a female assistant selects only 0.7pp more demographic information about a female worker relative to her average disclosure of 45.8% from Demographics in the case of a male worker ($p=0.16$; 1.5%). Table 3.B.7 illustrates that, compared to women, men provide significantly more information about female workers' marital status and number of children.

The assistants who are more likely to agree with traditional gender roles and stereotypes tend to disclose more demographic information about female workers relative to more tolerant assistants; this is captured by the marginally significant interaction term in column 2 of Table 3.B.6. This tendency suggests that stereotypes play a role in the differential treatment of women.

Gender of assistants and their stance toward traditional gender roles and stereotypes are correlated. Unsurprisingly, women score 0.130 more points on the index of tolerance toward women ($p<0.05$); while 59.7% of male assistants have a below-median tolerance index, only 48.6% of female assistants have a below-median tolerance index. To clarify the contributions of different subgroups, we report the results of heterogeneity analysis by assistants' gender and bias against women in Table 3.B.8. The higher disclosure rate of marital status of women is driven by male assistants, while female assistants do not contribute to this type of discrimination regardless of their tolerance index. On the other hand, both male and female assistants disclose more often the number of children for women, and the assistants biased against women tend to drive this effect more for both male and female assistants.

[16] Although we elicited the assistants' attitudes toward women after the main task (thus, after the treatment assignment), the index constructed—tolerance toward women—is balanced across treatment arms (see Table 3.B.4).

The effect of Female treatment on disclosure of demographic information is similar regardless of the profile quality (column 3 of Table 3.B.6). Therefore, women in various fields seem to face a similar treatment in this context.

3.4.2 Workers' gender and disclosure of work-related information

The negative effect of Female treatment on work-related information disclosure is driven especially by information about job responsibilities. Table 3.2 shows the results of regressions in which all information pieces from the Work section serve as dependent variables. The assistants are on average 7.2pp ($p<0.01$) and 1.8pp ($p<0.10$) less likely to disclose information about job responsibilities[17] and work area, respectively, if a CV has a female name (this corresponds to, respectively, a 12.6% and 2.5% decrease relative to the control means).

Table 3.2: Effect of a female worker's name on the disclosure of Work information

	(1) Status	(2) Area	(3) Position	(4) Experience	(5) Any responsibilities
Female	0.016	-0.018*	0.013	0.003	-0.072***
	(0.035)	(0.010)	(0.009)	(0.010)	(0.010)
Control mean	0.621	0.729	0.756	0.625	0.570
Observations	757	6056	6056	5299	6056

Notes: Regressions of specific work information pieces on the Female treatment indicator. *Status* is a binary variable equal to 1 if the assistant disclosed information that the worker is on parental leave (this information piece is present only in one profile). *Any responsibilities* is a binary variable equal to 1 if the assistant disclosed at least one job responsibility from the worker's profile. Regressions in Columns (2)-(5) include profile and assistant fixed effects. In these cases, the standard errors (in parentheses) are clustered at the assistant level. Column (1) is based on the OLS regression with the treatment indicator and assistants' age, gender, household size, educational and regional dummies, and recruitment experience (robust standard errors in parentheses). The control means are the average values of the outcomes in the male-CVs group.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The heterogeneity analysis by profile quality in column 3 of Table 3.B.9 reveals that

[17] In an alternative specification, we used the number of disclosed responsibilities as a dependent variable instead of a dummy indicating whether at least one job responsibility is disclosed (the workers' profiles include between 1 and 3 job responsibilities). We find that assistants select 0.12 fewer job responsibilities from CVs with female names ($p<0.01$; the average number of disclosed responsibilities from male CVs = 1.04).

the negative Female effect on work-related information disclosure is concentrated mainly among the low-quality profiles. Specifically, in high-quality profiles, the assistants disclose on average 0.5pp (p=0.51) less information about Work from female compared to male CVs (0.8% decrease); in low-quality profiles, disclosure of work-related information from female CVs is 3.6pp lower than from male CVs (p<0.01, 6.3% decrease). The main contributor to this lower disclosure from female low-quality CVs is the information about job responsibilities (column 4 of Table 3.B.10).

In line with the earlier finding that assistants provide more family-related information about female workers, differences in disclosure of the work-related information may also be connected to gender stereotypes. This link is supported by (i) the finding that assistants are 3.5pp more likely to provide information about job positions from female low-quality CVs than from male low-quality CVs (column 2 of Table 3.B.10, p<0.05)[18] and (ii) the observation that our low-quality profiles tend to feature female-dominated occupations (e.g., cashier, postal delivery, administrative worker). To show more explicitly that assistants tend to over-provide stereotypical information about female-workers' jobs, we run heterogeneity analyses by female- vs. male-dominated occupations in Table 3.B.11. The positive Female effect on the disclosure of a job position is clearly concentrated among the profiles with female-dominated occupations.[19] Additional heterogeneity analysis by assistants' gender reveals that male assistants are 6.4pp more likely to provide information about job positions from female low-quality CVs (Table 3.B.12, p<0.01, 7.9% increase relative to their mean disclosure from male low-quality CVs). In comparison, female assistants are only 0.5pp more likely to disclose information about job positions from female low-quality CVs (p=0.81, 0.6% increase relative to their mean disclosure from male low-quality CVs).[20] Taken together with the earlier observation that men select more family-related information about female workers, this result suggests that gender

[18]There are no such differences in the case of high-quality CVs.

[19]We classified profiles 1, 5, 7 as female-dominated, and 2, 6, 8 as male-dominated; profiles 3 and 4 are ambiguous, so we excluded them from this analysis. We also ran the same heterogeneity analyses restricted further to profiles with even more obvious classification as female- or male-dominated occupations and the results hold, although they lose significance in the most restrictive specification due to the substantial sample reduction (these analyses are available upon request).

[20]We continued to split the sample by profile quality instead of gender-dominated occupations because this is our pre-specified heterogeneity analysis. The assistants' gender differences are confirmed by the specification that classifies profiles into "female-" and "male-dominated" groups.

stereotypes play a more prominent role in the selection of information by male assistants.

We conclude this section by commenting on attention outcomes. There are no significant effects of a female name on attention outcomes, but there seems to be a tendency to lower attention to female CVs (see column 1 of Table 3.B.13 for the time spent on a CV and column 2 of Table 3.B.13 for the clicks on the "learn-more" buttons). However, we do not have data on assistants' attention to *all* individual information pieces because we only recorded the time that the assistants spent on the entire CV, and the "learn-more" buttons were presented only next to pieces that were likely to require additional explanation. Therefore, we leave to future work the investigation of the attentional underpinning of discrimination in the disclosure of the specific information pieces that we identified.

3.5 Conclusion

We use a novel experimental design to study discrimination in information transmission in the context of hiring. We create an online labor market in which our main subjects, respondents who act as human resource assistants, select information about workers for other respondents, who act as hiring managers. The managers inspect only the information selected for them and make hiring decisions about the workers. The exogenous variation in our experiment comes from random names that we assign to the workers' profiles to signal gender.

Our results indicate that assistants tend to disclose information differently depending on the gender of the workers. First, we document that assistants provide more information about family and less information about work from female CVs. A closer look at the pieces of information disclosed suggests that differential disclosure is driven by gender stereotypes. In particular, the selection from female CVs is more likely to contain information about the marital status, number of children, and female-dominated occupation than the selection from male CVs.

Our findings have several practical implications. First, HR assistants may discriminate unintentionally, and thus simply informing them about our findings may induce them to rethink their practices and adjust their training programs. Second, our research invites the design of more discrimination-proof communication protocols. Although some

businesses are already using standardized hiring processes with prescribed rules, our discussions with human resource professionals suggest that this is not always the case and that there is room for (more subtle) differential communication about different groups of candidates. Finally, the emphasis that our assistants put on family-related information for females suggests the importance of a more general societal problem related to childcare and unequal gender roles. Among other things, this calls for expansion of affordable childcare availability and parental leave programs that minimize the (perceived) loss of firms related to childcare and that promote shared parental leave between fathers and mothers.

Our work can serve as a motivation for investigating what other channels, similar to differential disclosure, may underlie biases in hiring of female applicants. For example, using gendered language in job-position descriptions that emphasizes masculine-associated traits as desired qualities may discourage many talented women from applying.

3.A Appendix figures

Figure 3.A.1: Overview of the project

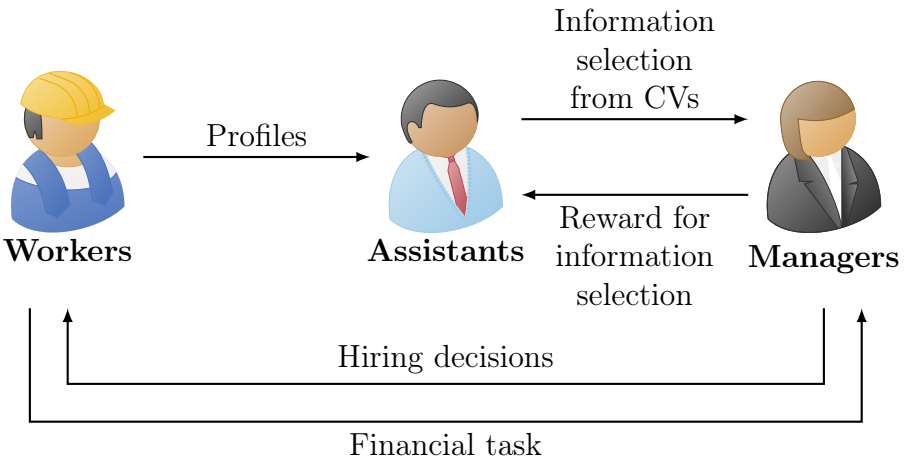
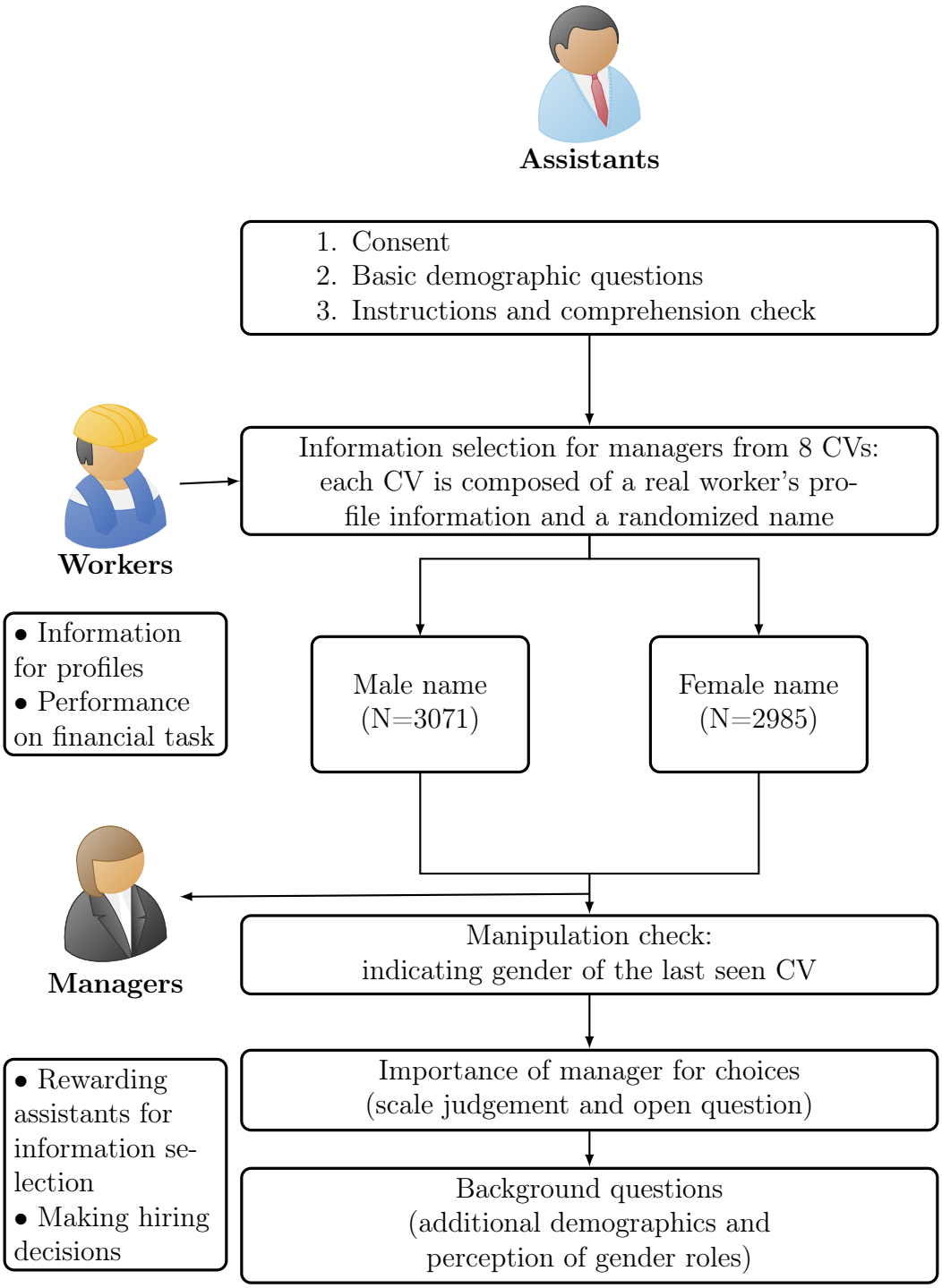


Figure 3.A.2: Flow of the experiment with assistants and the connections to the surveys with workers and managers



3.B Appendix tables

Table 3.B.1: Demographic composition of our sample of assistants compared to the general Czech population

	Mean: experiment (assistants)	Mean: Demographic Yearbook of the Czech Republic 2020
Gender		
Male	0.51	0.51
Female	0.49	0.49
Age group		
18 to 24 years	0.103	0.102
25 to 34 years	0.211	0.209
35 to 44 years	0.255	0.257
45 to 54 years	0.233	0.233
55 to 64 years	0.198	0.199
Education		
Primary and secondary without national school-leaving exam	0.414	0.417
Secondary with national school-leaving exam	0.375	0.373
University	0.211	0.210
Region (NUTS 2)		
Prague	0.127	0.127
Central Bohemia	0.130	0.129
Southwest	0.114	0.115
Northwest	0.104	0.104
Northeast	0.141	0.140
Southeast	0.156	0.159
Central Moravia	0.116	0.113
Moravian-Silesian	0.112	0.113

Notes: This table compares the shares of selected socio-demographic groups in our experiment (N=757) to the corresponding shares received from the Demographic Yearbook of the Czech Republic 2020.

Table 3.B.2: Summary statistics for assistants' sample

	(1) mean	(2) sd	(3) p50	(4) min	(5) max
Female	0.49	0.50	0.00	0.00	1.00
Age	42.04	12.92	41.00	18.00	64.00
Household size	2.76	1.19	3.00	1.00	6.00
Primary and secondary education without national school-leaving exam	0.41	0.49	0.00	0.00	1.00
Secondary education with national school-leaving exam	0.38	0.48	0.00	0.00	1.00
University degree	0.21	0.41	0.00	0.00	1.00
Prague	0.13	0.33	0.00	0.00	1.00
Central Bohemia	0.13	0.34	0.00	0.00	1.00
Southwest	0.11	0.32	0.00	0.00	1.00
Northwest	0.10	0.31	0.00	0.00	1.00
Northeast	0.14	0.35	0.00	0.00	1.00
Southeast	0.16	0.36	0.00	0.00	1.00
Central Moravia	0.12	0.32	0.00	0.00	1.00
Moravian Silesian	0.11	0.32	0.00	0.00	1.00
Employed	0.75	0.43	1.00	0.00	1.00
Unemployed	0.02	0.14	0.00	0.00	1.00
Household net monthly income > CZK 50,000	0.31	0.46	0.00	0.00	1.00
Has recruitment experience	0.13	0.33	0.00	0.00	1.00
Thought about the hiring manager	8.15	1.98	9.00	0.00	10.00
Correctly identified last worker's gender	0.92	0.27	1.00	0.00	1.00

Notes: This table presents the summary statistics for 757 assistants. 6 assistants (< 1%) and 72 assistants (9.5%) did not record their employment status and income, respectively. We chose CZK 50,000 as a threshold value for income because net monthly household income of a median subject lies between CZK 40,001 and 50,000. An assistant could select whether he/she did not think about the manager at all (0 on a numeric scale) or a lot (10 on a numeric scale) while selecting information about workers. According to the Czech Statistical Office (2021), the share of employed people in the total Czech population aged 15-64 years was 75.1% and the unemployment rate in the same age group was 2.2% in December 2021 (we did not find corresponding statistics for the group aged 18-64 years, which would be the same age range as our sample of assistants covers). The unemployment rate in our sample, calculated by dividing the number of unemployed participants by the sum of employed and unemployed individuals, is equal to 2.7%. The net monthly household income of a median assistant lies between 40,001 and 50,000 Czech crowns (the dollar equivalents are approximately $1,690 and $2,110, respectively), which is somewhat higher than the statistics based on the data from the Czech Statistical Office (37,436 Czech crowns in 2021—to calculate this number, we multiplied monthly net income per capita of a median household by the average number of the median household members; the inputs were obtained from Table 2a at https://www.czso.cz/csu/czso/household-income-and-living-conditions-6yp06pfzwa).

Table 3.B.3: List of workers' names used in the experiment

Profile	Name	Gender
1	**PETR**	Male
	VOLODYMYR	Male
	ADÉLA	Female
	OLEKSANDRA	Female
2	**ONDŘEJ**	Male
	EVGENIY	Male
	KATEŘINA	Female
	YEKATERINA	Female
3	**JINDŘICH**	Male
	MYKHAILO	Male
	MARKÉTA	Female
	OLESYA	Female
4	**VOJTĚCH**	Male
	YURIY	Male
	ZDEŇKA	Female
	VASILISA	Female
5	**MATĚJ**	Male
	DMITRIY	Male
	LUCIE	Female
	KSENIYA	Female
6	**JIŘÍ**	Male
	OLEXIY	Male
	JITKA	Female
	OLENA	Female
7	**ZDENĚK**	Male
	VASILY	Male
	ALŽBĚTA	Female
	YELYZAVETA	Female
8	**RADEK**	Male
	ANATOLIY	Male
	BOŽENA	Female
	VARVARA	Female

Notes: This table shows the list of workers' names used in our experiment. Each profile had two female and two male names, each version having one local and one foreign-sounding name.

Table 3.B.4: Randomization check I (assistants)

	(1) Male (control)	(2) Female (treatment)	(3) t-test	(4) N
Female	0.49	0.49	0.76	6056
Age	42.23	41.83	0.22	6056
Household size	2.73	2.79	0.05	6056
Primary and secondary education without national school-leaving exam	0.41	0.42	0.76	6056
Secondary education with national school-leaving exam	0.38	0.37	0.17	6056
University degree	0.20	0.22	0.21	6056
Prague	0.13	0.12	0.56	6056
Central Bohemia	0.13	0.13	0.73	6056
Southwest	0.11	0.11	0.82	6056
Northwest	0.10	0.11	0.12	6056
Southeast	0.15	0.16	0.27	6056
Northeast	0.14	0.14	0.82	6056
Central Moravia	0.12	0.11	0.11	6056
Moravia-Silesia	0.12	0.11	0.16	6056
Employed	0.75	0.75	0.84	6008
Unemployed	0.02	0.02	0.82	6008
Income is missing	0.10	0.09	0.60	6056
Household net monthly income $>$ CZK 50,000	0.29	0.33	0.01	5480
Has recruitment experience (dummy)	0.13	0.12	0.07	6056
Thought about the hiring manager	8.15	8.14	0.79	6056
Correctly identified last worker's gender	0.92	0.91	0.32	6056
Tolerance to women	3.26	3.25	0.61	6056
Mobile survey completion	0.40	0.42	0.36	6056
N	3071	2985		

Notes: Means of assistants' characteristics in the control and treatment group. Column (3) reports p-values of t-test for the hypothesis that the means are equal in the two groups. The tolerance index was constructed by taking averages of responses to 7 questions regarding women (all measured on a scale from 1 to 5; when necessary, we recoded responses so that 5 would mean the highest tolerance).

Table 3.B.5: Randomization check II (assistants)

	(1) Female treatment
Female	-0.005
	(0.013)
Age	-0.000
	(0.001)
Household size	0.009
	(0.006)
Primary and secondary	-0.022
education without national school-leaving exam	(0.018)
Secondary education with national school-leaving exam	-0.033*
	(0.018)
Central Bohemia	0.008
	(0.026)
Southwest	0.009
	(0.027)
Northwest	0.039
	(0.027)
Southeast	0.025
	(0.025)
Northeast	0.010
	(0.025)
Central Moravia	-0.026
	(0.026)
Moravian Silesian	-0.017
	(0.027)
Income is missing	-0.015
	(0.022)
Has recruitment experience	-0.040**
	(0.020)
Tolerance to women	-0.004
	(0.009)
Correctly identified last worker's gender	-0.030
	(0.024)
Mobile survey completion	0.006
	(0.014)
Thought about the hiring manager	-0.000
	(0.003)
Constant	0.545***
	(0.057)
N	6056
F	1.180
p-value of F-test	0.268

Notes: Regression of the treatment indicator on assistants' characteristics. Standard errors in parentheses. We include only covariates which do not have missings because our treatment effects are estimated on the full sample. If we include the high-income and employed dummies, the p-value of F-test is 0.217.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.6: Heterogeneity analyses for the effect of a female worker's name on the share of disclosed Demographic information

	Share of disclosed pieces (Demographics)		
	(1)	(2)	(3)
Female (a)	0.032***	0.013**	0.018***
	(0.006)	(0.006)	(0.005)
Female * Female assistant (b)	-0.025***		
	(0.008)		
Female * Biased against women (c)		0.013*	
		(0.008)	
Female * Low-quality profile (d)			0.005
			(0.007)
(a) + (b)	0.007		
	(0.005)		
(a) + (c)		0.026***	
		(0.005)	
(a) + (d)			0.023***
			(0.005)
Control mean	0.494	0.455	0.477
N	6056	6056	6056

Notes: All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. *Female* is a treatment indicator equal to 1 if a CV has a female name. *Biased against women* is equal to 1 if an index of tolerance toward women is less or equal to its median value (see Section 3.4.1 for the details about the construction of the tolerance index). The control means are the average values of the outcome in the male-CVs group and: in Column (1), a subsample of male assistants; in Column (2), a subsample of "tolerant" assistants; in Column (3), a subsample of high-quality profiles.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.7: Heterogeneity analysis by assistant's gender for the effect of a female worker's name on the disclosed Demographic information

	(1) Age	(2) Marital status	(3) Number of children	(4) Driving license	(5) Surveys
Female (a)	-0.005	0.055***	0.103***	0.001	0.004
	(0.007)	(0.012)	(0.015)	(0.008)	(0.011)
Female * Female assistant (b)	-0.001	-0.064***	-0.044**	-0.009	-0.008
	(0.011)	(0.016)	(0.020)	(0.011)	(0.014)
(a) + (b)	-0.005	-0.009	0.059***	-0.007	-0.004
	(0.008)	(0.011)	(0.013)	(0.008)	(0.009)
Control mean	0.781	0.423	0.292	0.717	0.242
N	6056	5299	6056	6056	6056

Notes: All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. *Female* is a treatment indicator equal to 1 if a CV has a female name. The control means are the average values of the outcomes in the male-CVs group and subsample of male assistants. *Surveys* informs about the actual number of surveys that a worker completed in the past.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.8: Heterogeneity analyses for the effect of female worker's name on disclosed Demographic information by assistants' gender and bias

	(1) Age	(2) Marital status	(3) Number of children	(4) Driving license	(5) Surveys
Female (a)	-0.003	0.039**	0.086***	0.008	0.001
	(0.012)	(0.016)	(0.021)	(0.014)	(0.017)
Female * Female assistant (b)	-0.013	-0.047**	-0.045*	-0.008	-0.020
	(0.015)	(0.023)	(0.027)	(0.017)	(0.021)
Female * Biased against women (c)	-0.003	0.028	0.028	-0.012	0.006
	(0.015)	(0.024)	(0.029)	(0.017)	(0.022)
Female * Female assistant * Biased against women (d)	0.025	-0.029	0.010	-0.005	0.026
	(0.023)	(0.032)	(0.039)	(0.023)	(0.028)
(a) + (b)	-0.016*	-0.009	0.041**	0.001	-0.019
	(0.009)	(0.016)	(0.017)	(0.010)	(0.012)
(a) + (c)	-0.006	0.067***	0.114***	-0.004	0.007
	(0.009)	(0.017)	(0.020)	(0.010)	(0.013)
(a) + (b) + (c) + (d)	0.006	-0.010	0.078***	-0.016	0.013
	(0.014)	(0.013)	(0.019)	(0.011)	(0.012)
Control mean	0.749	0.419	0.273	0.734	0.220
N	6056	5299	6056	6056	6056

Notes: All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. *Female* is a treatment indicator equal to 1 if a CV has a female name. *Surveys* informs about the actual number of surveys that a worker completed in the past. The control means are the average values of the outcomes in the male-CVs group and subsample of male assistants with above median tolerance index (indicating "no bias against women").

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.9: Heterogeneity analyses for the effect of a female worker's name on the share of disclosed Work information

	Share of disclosed pieces (Work)		
	(1)	(2)	(3)
Female (a)	-0.012*	-0.021***	-0.005
	(0.007)	(0.007)	(0.007)
Female * Female assistant (b)	-0.017*		
	(0.009)		
Female * Biased against women (c)		0.002	
		(0.009)	
Female * Low-quality profile (d)			-0.031***
(a) + (b)	-0.029***		
	(0.007)		
(a) + (c)		-0.019***	
		(0.007)	
(a) + (d)			-0.036***
			(0.007)
Control mean	0.588	0.606	0.606
N	6056	6056	6056

Notes: All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. *Female* is a treatment indicator equal to 1 if a CV has a female name. *Biased against women* is equal to 1 if an index of tolerance toward women is less or equal to its median value (see Section 3.4.1 for the details about the construction of the tolerance index). The control means are the average values of the outcome in the male-CVs group and: in Column (1), a subsample of male assistants; in Column (2), a subsample of tolerant assistants; in Column (3), a subsample of high-quality profiles.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.10: Heterogeneity analyses for the effect of a female worker's name on disclosed Work information by profile quality

	(1) Area	(2) Position	(3) Experience	(4) Any responsibilities
Female (a)	-0.021	-0.009	0.018	-0.002
	(0.013)	(0.012)	(0.014)	(0.014)
Female * Low-quality profile (b)	0.006	0.043**	-0.025	-0.140***
	(0.018)	(0.018)	(0.020)	(0.021)
(a) + (b)	-0.015	0.035**	-0.007	-0.142***
	(0.013)	(0.013)	(0.014)	(0.015)
Control mean	0.774	0.808	0.638	0.642
N	6056	6056	5299	6056

Notes: All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. *Female* is a treatment indicator equal to 1 if a CV has a female name. *Any responsibilities* is a binary variable equal to 1 if an assistant disclosed at least one job responsibility from a CV. The control means are the average values of the outcomes in the male-CVs group and subsample of high-quality profiles.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.11: Heterogeneity analysis by female-dominated occupation for the effect of a female worker's name on disclosed Work information

	(1) Area	(2) Position	(3) Experience	(4) Any responsibilities
Female (a)	-0.012	-0.022	0.018	-0.009
	(0.015)	(0.014)	(0.015)	(0.017)
Female * Female-dominated job (b)	-0.010	0.055**	-0.023	-0.173***
	(0.022)	(0.022)	(0.022)	(0.026)
(a) + (b)	-0.021	0.034**	-0.005	-0.182***
	(0.016)	(0.016)	(0.016)	(0.018)
Control mean	0.765	0.831	0.638	0.633
N	4542	4542	4542	4542

Notes: All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. *Female* is a treatment indicator equal to 1 if a CV has a female name. *Female-dominated job* is a binary variable equal to 1 if a CV is based on a profile with a female-dominated occupation (we classify profiles 1, 5, 7 as female-dominated, and 2, 6, 8 as male-dominated; profiles 3 and 4 are ambiguous so we exclude them from this analysis). *Any responsibilities* is a binary variable equal to 1 if an assistant disclosed at least one job responsibility from a CV. The control means are the average values of the outcomes in the subsample of CVs with male names and male-dominated occupations.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.12: Heterogeneity analyses for the effect of a female worker's name on disclosed Work information by profile quality and assistants' gender

	(1) Area	(2) Position	(3) Experience	(4) Any responsibilities
Female (a)	-0.028	-0.008	0.034*	0.037*
	(0.019)	(0.018)	(0.021)	(0.020)
Female * Female assistant (b)	0.014	-0.003	-0.033	-0.080***
	(0.026)	(0.023)	(0.028)	(0.028)
Female * Low-quality profile (c)	0.011	0.072***	-0.029	-0.178***
	(0.026)	(0.026)	(0.029)	(0.030)
Female assistant * Low-quality profile	-0.008	0.029	-0.014	-0.020
	(0.025)	(0.026)	(0.027)	(0.030)
Female * Female assistant * Low-quality profile (d)	-0.011	-0.056	0.008	0.077*
	(0.035)	(0.036)	(0.039)	(0.041)
(a) + (b)	-0.014	-0.011	0.001	-0.043**
	(0.017)	(0.015)	(0.020)	(0.019)
(a) + (c)	-0.017	0.064***	0.005	-0.141***
	(0.019)	(0.019)	(0.021)	(0.021)
(a) + (b) + (c) + (d)	-0.013	0.005	-0.020	-0.144***
	(0.019)	(0.019)	(0.018)	(0.020)
Control mean	0.778	0.808	0.615	0.643
N	6056	6056	5299	6056

Notes: All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. *Female* is a treatment indicator equal to 1 if a CV has a female name. *Any responsibilities* is a binary variable equal to 1 if an assistant disclosed at least one job responsibility from a CV. The control means are the average values of the outcomes in the male-CVs group and subsample of male assistants and high-quality profiles.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.13: Effect of female name on attention measures

	(1) Time	(2) Learn-more clicks
Female	-32.869	-0.018
	(32.814)	(0.040)
Control mean	129.232	0.693
Observations	6056	6056

Notes: All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. *Female* is a treatment indicator equal to 1 if a CV has a female name. *Time* is the number of seconds that an assistant spent on selecting information from a CV. *Learn-more clicks* is the number of clicks that an assistant made on "More information" buttons embedded in a CV. The control means are the average values of the outcomes in the male-CVs group.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.14: Effect of female worker's name on the share of disclosed information in the subsample of assistants with recruitment experience

	(1) Overall	(2) Summary	(3) Demographics	(4) Education	(5) Work	(6) Certificates	(7) Skills	(8) Interests	(9) Volunteering
Female	-0.003	-0.001	0.001	-0.001	-0.005	0.013	-0.016	0.006	0.021
	(0.007)	(0.014)	(0.010)	(0.014)	(0.015)	(0.015)	(0.012)	(0.015)	(0.026)
Control mean	0.524	0.576	0.502	0.548	0.569	0.510	0.576	0.352	0.353
Observations	760	760	760	760	760	760	760	760	380

Notes: All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. *Female* is a treatment indicator equal to 1 if a CV has a female name. The control means are the average values of the outcomes in the male-CVs group.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.15: Effect of female worker's name on the disclosuse of Demographic information in the subsample of assistants with recruitment experience

	(1) Age	(2) Marital status	(3) Children	(4) Driving license	(5) Surveys
Female	-0.029**	-0.009	0.051*	-0.033**	0.002
	(0.014)	(0.022)	(0.028)	(0.014)	(0.020)
Control mean	0.758	0.435	0.303	0.726	0.274
Observations	760	665	760	760	760

Note: Regressions of specific demographic information pieces on the Female treatment indicator. *Surveys* contains the actual number of surveys that a worker completed in the past. All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. The control means are the average values of the outcomes in the male-CVs group.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.16: Effect of female worker's name on the disclosuse of Work information in the subsample of assistants with recruitment experience

	(1) Area	(2) Position	(3) Experience	(4) Any responsibilities
Female	0.020	0.021	-0.021	-0.058**
	(0.029)	(0.022)	(0.030)	(0.024)
Control mean	0.675	0.692	0.609	0.572
Observations	760	760	665	760

Notes: Regressions of specific work information pieces on the Female treatment indicator. *Any responsibilities* is a binary variable equal to 1 if the assistant disclosed at least one job responsibility from the worker's profile. All regressions include profile and assistant fixed effects. Standard errors (in parentheses) are clustered at the assistant level. The control means are the average values of the outcomes in the male-CVs group.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.C Assistants' instructions (translated from Czech)

Hello,

Participation in this survey is totally voluntary. If you start the survey and you no longer wish to finish it, you can do so without any consequences.

If you decide to participate in the survey, make sure that you have enough time to finish it (i.e., at least **25 minutes**), please.

For completion of the survey, you will receive the **reward stated in the invitation**. In addition, you may receive a **bonus** whose amount depends partially on your decisions. You will receive the bonus points in February 2022 at latest, after the evaluation of the whole survey.

In contrast to traditional survey questions, which are about hypothetical situations, **you will now make decisions that might have real** (financial) **consequences for other participants of our online labor market**. Specifically, you will select information from profiles of workers.

We would like to assure you that **panel iVýzkumy.cz guarantees your total anonymity and the confidentiality of your answers.**

Please answer the questions truthfully, according to your own judgement and knowledge, regardless of whether your opinions adhere to mainstream attitudes or are politically correct. It is crucial for success of the survey that you go attentively through the whole survey and adhere to the instructions in each part of the survey.

If you are done reading the text above and agree to participate in this survey, please check "Yes". You will start the survey by pressing the button →.

◯ Yes
◯ No

[Next page]

What is your **gender**?

◯ Man

◯ Woman

What is your **age**?

Enter a number into the following field:

What is your **highest completed education**?

◯ Unfinished elementary

◯ Elementary

◯ Vocational or general secondary without state examination

◯ Secondary with state examination

◯ Higher professional

◯ University

In what **region** do you reside?

We want to know the region where you actually live, not the region of your permanent residency. Click on the arrow below to show the list of regions.

∨

How many **people** are there in **your household** (including you)?

◯ 1

◯ 2

◯ 3

◯ 4

◯ 5

◯ More, write how many:

[Next page]

In this survey, you will act as an **assistant in hiring workers** in our **online labor market**.

We emphasize that, in contrast to traditional survey questions, which are about hypothetical situations, you will now make **decisions** that might have **real** (financial) **consequences** for other participants of our survey.

[Next page]

Your task will be to **review 8** workers' **profiles** and **select** only those **pieces of information** that you would like to **provide** to another Czech participant of our survey – this person will act as a **hiring manager**.

The **manager** will **hire** workers for a **financial task**, which consists of a **series of various financial decisions**, e.g. about investments.

The **manager** will be deciding about each of these 8 people individually, i.e. he/she might **hire any number of people** (e.g. all 8 or even nobody).

[Next page]

The **manager** will be **busy** because he/she will have to make multiple hiring decisions during a limited time. Therefore, **your task** of **simplifying the profiles** is a crucial help to him/her.

Before making a hiring decision about a worker, the **manager** will see **only the information** that **you will select**, but he/she will never see the workers' original profiles.

Some pieces of information will have a **button** More information next to them that will enable you to better understand the corresponding piece of information, but this button **will be never displayed to the manager**. Hence, if you choose the corresponding piece of information, the manager will see only its content, but not the button with the additional information.

[Next page]

Each profile simplification may be important because **your information selection** might **impact** the manager's **decisions** and bear **financial consequences**.

Workers who are **hired will receive extra money**. The **manager** will receive a **higher reward** if the hired workers perform well on the financial task.

[Next page]

It is important that you select information for the **manager** diligently because he/she **will decide how to reward your effort**. This **reward** will be paid **in addition** to your participation fee.

If the **manager** finds **your information selection useful**, he/she can give you **up to 500 points**, which costs him/her nothing. If the manager finds that your information selection is not useful at all, he/she might give you 0 points.
[Note: participants were rewarded by the data collection agency's points with a conversion rate 10 points = 1 CZK.]

During the survey, you will be able to return to these instructions.

[Next page]

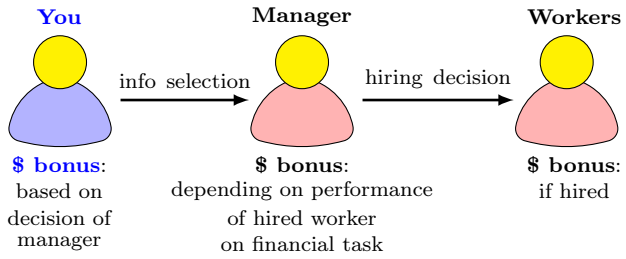
In this part, we would like to **check** your **understanding** of the task instructions that you just read. If you want to go through the **instructions one more time**, press the button ←.

For each of the following statements, please decide whether it is true or false.

	Yes	No
I will see profiles of **8 workers**. **My task** is to **select information** from these profiles for another Czech participant who will act as a **hiring manager**.	◯	◯
The manager is **hiring** people for a **financial task**. Besides information that I select, the **manager will NOT see the full profiles**. The **manager** will be **busy** making many hiring decisions.	◯	◯
The manager will determine **my bonus** according to how **useful** he/she finds **my information selection**.	◯	◯

[Next page]

All statements on the previous page were **CORRECT**.



[next page]

HERE you can recheck the instructions (they will open in a new tab).

Now you will look through a worker's profile. Please **select** the pieces of **information** that you would like to **provide to a manager** who will consider this worker for the **task** that consists of a **series of financial decisions**.

ONDŘEJ (ID 664)[21]

SUMMARY (based on self-evaluation)

- ☐ Ondřej has logical and technically oriented mindset. He behaves consistently and is eager to learn.
- ☐ OndŘej is sometimes inattentive to details and lacks self-confidence.
- ☐ According to Ondřej it is quite likely that [he/she] would be able to convince other people of [his/her] opinion in financial services.

DEMOGRAPHICS

- ☐ **Age:** 27
- ☐ **Marital status:** single
- ☐ **Number of children:** 0
- ☐ **Driving license:** yes

[21] To authentically illustrate the questionnaire of assistants, we present the 8 profiles in a random order and randomly select one of the corresponding names for each profile from Table 3.B.3. The presented order of profiles in this Appendix is: 2, 5, 7, 6, 1, 4, 8, 3. Note, however, that each assistant could see the profiles in a different order and with different names than displayed in this Appendix.

☐ **Number of completed online questionnaires:** 15 More information

[After clicking on "More information":]

The number of completed online questionnaires is a record from a database about the actual total number of online surveys that the worker has properly completed in the past. Note: This button with more information will never be displayed to the manager.

EDUCATION

☐ **Level:** university – master's degree

☐ **Area of studies:** economics

☐ **Favorite subject:** risk management More information

[After clicking on "More information":]

Risk management is an area of managing projects and processes that deals with determination and evaluation of their risks and undesirable effects. Note: This button with more information will never be displayed to the manager.

☐ **Favorite subject:** mathematics

☐ **Favorite subject:** financial analysis

WORK

☐ **Employment sector:** banking

☐ **Current position:** analyst

☐ **Work experience in the current position:** 2 years

☐ **Job responsibilities:** error analysis

☐ **Job responsibilities:** preparation of reports

☐ **Job responsibilities:** accounting control

CERTIFICATES (based on real tasks)

☐ Attained **6 points in a math test** (**average** of all candidates: **4** points) More information

[After clicking on "More information":]

The Math test included 10 questions that tested the knowledge of basic Math operations, equation solving, etc. Participants had 2.5 minutes to complete the test. Note: This button with more information will never be displayed to the manager.

☐ Attained **5 points in a financial-literacy quiz** (**average** of all candidates: **3.9** points) More information

[After clicking on "More information":]

The financial-literacy quiz included 5 questions that tested the understanding of basic financial concepts (inflation, interest, etc.). Note: This button with more information will never be displayed to the manager.

☐ Attained **6 points in a slider task** (**average** of all candidates: **24** points) More information

[After clicking on "More information":]

The slider task is a mechanical task in which participants had to center within a 2-minute limit as many sliders as possible (max. 48) with a random initial position. Note: This button with more information will never be displayed to the manager.

VOLUNTEERING (based on real decisions)

☐ Completed a survey for free in order to **donate the money to a charity** More information

[After clicking on "More information":]

In the questionnaire, we asked the worker whether he or she is willing to participate in another survey in upcoming days and donate the reward from participation to a charity of own choice. If the worker agreed, he or she received later an invitation for a survey in which it was explicitly mentioned that the reward will be donated. We could verify whether the worker truly completed this survey. Note: This button with more information will never be displayed to the manager.

☐ **Donated own participation fee** in **56%** of completed online surveys

SKILLS

☐ **Microsoft Word:** advanced

☐ **Microsoft Excel:** advanced

☐ **Microsoft PowerPoint:** advanced

☐ **Internet banking:** using

☐ **English language:** good knowledge

☐ **Has experience with** data analysis

☐ **Has experience with** economics

☐ **Has experience with** data entry

INTERESTS

☐ Sport activities

☐ Traveling

☐ International news

[next page]

Based on **your** earlier **selection**, the hiring **manager will see** the following information:
***

ONDŘEJ (ID 664)

[At this point, the assistant saw what would be displayed to the manager about this worker based on the assistant's choices. This means also that the assistant could see that the More information buttons were going to be suppressed.]

***

If you want to return to the profile of the worker and **change your selection of information, press** button ←.

[next page]

HERE you can recheck the instructions (they will open in a new tab).

LUCIE (ID 141)

Now you will look through a worker's profile. Please **select** the pieces of **information** that you would like to **provide to a manager** who will consider this worker for the **task** that consists of a **series of financial decisions**.

SUMMARY (based on self-evaluation)

☐ Lucie is stress-resistant. Her strengths are credibility and responsibility.

☐ Lucie sometimes postpones things and is inattentive to details.

DEMOGRAPHICS

☐ **Age:** 31

☐ **Number of children:** 1

☐ **Driving license:** yes

☐ **Number of completed online questionnaires:** 15 More information

[After clicking on "More information":]

The number of completed online questionnaires is a record from a database about the actual total number of online surveys that the worker has properly completed in the past. Note: This button with more information will never be displayed to the manager.

EDUCATION

☐ **Level:** secondary (without school leaving exam)

☐ **Area of studies:** storage operator More information

[After clicking on "More information":]

A storage operator is primarily responsible for logistic operations with physical products: receipt of materials in a warehouse, management of warehouse records and administration, handling of materials, packing, and preparation of goods for expedition. Note: This button with more information will never be displayed to the manager.

WORK

☐ **Current work status:** on [maternal/parental] leave

☐ **Last employer:** post office

☐ **Last position:** delivery

☐ **Work experience in the last position:** 3 years

☐ **Job responsibilities:** communication with people

CERTIFICATES (based on real tasks)

- ☐ Attained **1 point in a math test** (**average** of all candidates: **4** points) More information
  [After clicking on "More information":]
  The Math test included 10 questions that tested the knowledge of basic Math operations, equation solving, etc. Participants had 2.5 minutes to complete the test. Note: This button with more information will never be displayed to the manager.

- ☐ Attained **2 points in a financial-literacy quiz** (**average** of all candidates: **3.9** points) More information
  [After clicking on "More information":]
  The financial-literacy quiz included 5 questions that tested the understanding of basic financial concepts (inflation, interest, etc.). Note: This button with more information will never be displayed to the manager.

- ☐ Attained **18 points in a slider task** (**average** of all candidates: **24** points) More information
  [After clicking on "More information":]
  The slider task is a mechanical task in which participants had to center within a 2-minute limit as many sliders as possible (max. 48) with a random initial position. Note: This button with more information will never be displayed to the manager.

SKILLS

- ☐ **Microsoft Word:** basic knowledge
- ☐ **Microsoft Excel:** basic knowledge
- ☐ **Microsoft PowerPoint:** basic knowledge
- ☐ **Internet banking:** not using
- ☐ **English language:** partial knowledge
- ☐ **Has experience with** data entry

INTERESTS

- ☐ Watching TV
- ☐ Sometimes reads Blesk [Blesk is a Czech tabloid]

[next page]

Based on **your** earlier **selection**, the hiring **manager will see** the following information:
***

LUCIE (ID 141)

[At this point, the assistant saw what would be displayed to the manager about this worker based on the assistant's choices.]

***

If you want to return to the profile of the worker and **change your selection of information, press** button ←.

[next page]

HERE you can recheck the instructions (they will open in a new tab).

Now you will look through a worker's profile. Please **select** the pieces of **information** that you would like to **provide to a manager** who will consider this worker for the **task** that consists of a **series of financial decisions**.

ЕЛИЗАВЕТА (YELYZAVETA) (ID 812)

SUMMARY (based on self-evaluation)

- ☐ Yelyzaveta is efficient. She is responsible and able to solve difficult and complex problems.
- ☐ Yelyzaveta sometimes postpones things. She is impulsive and bad at financial management.

DEMOGRAPHICS

- ☐ **Age:** 38
- ☐ **Marital status:** married
- ☐ **Number of children:** 1
- ☐ **Driving license:** yes
- ☐ **Number of completed online questionnaires:** 7 More information

[After clicking on "More information":]
The number of completed online questionnaires is a record from a database about the actual total number of online surveys that the worker has properly completed in the past. Note: This button with more information will never be displayed to the manager.

EDUCATION

- ☐ **Level:** secondary (without school leaving exam)
- ☐ **Area of studies:** administration
- ☐ **Favorite subject:** theory
- ☐ **Favorite subject:** practice

WORK

- ☐ **Employment sector:** trucking
- ☐ **Current position:** administrative worker
- ☐ **Work experience in the last position:** 6 years
- ☐ **Job responsibilities:** paperwork

CERTIFICATES (based on real tasks)

- ☐ Attained **3 points in a math test (average** of all candidates: **4** points) More information
  [After clicking on "More information":]
  The Math test included 10 questions that tested the knowledge of basic Math operations, equation solving, etc. Participants had 2.5 minutes to complete the test. Note: This button with more information will never be displayed to the manager.
- ☐ Attained **4 points in a financial-literacy quiz (average** of all candidates: **3.9** points) More information
  [After clicking on "More information":]
  The financial-literacy quiz included 5 questions that tested the understanding of basic financial concepts (inflation, interest, etc.). Note: This button with more information will never be displayed to the manager.

☐ Attained **44 points in a slider task (average** of all candidates: **24** points)
More information

[After clicking on "More information":]

The slider task is a mechanical task in which participants had to center within a 2-minute limit as many sliders as possible (max. 48) with a random initial position.

Note: This button with more information will never be displayed to the manager.

VOLUNTEERING (based on real decisions)

☐ **Donated own participation fee** in **100%** of completed online surveys

SKILLS

☐ **Microsoft Word:** professional

☐ **Microsoft Excel:** professional

☐ **Microsoft PowerPoint:** professional

☐ **Internet banking:** not using

☐ **English language:** partial knowledge

☐ **Has experience with** administrative work

☐ **Has experience with** building savings

INTERESTS

☐ Reading books

☐ Cooking

☐ Sport activities

☐ Reading business literature

[next page]

Based on **your** earlier **selection**, the hiring **manager will see** the following information:

***

ЕЛИЗАВЕТА (YELYZAVETA) (ID 812)

[At this point, the assistant saw what would be displayed to the manager about this worker based on the assistant's choices.]

***

If you want to return to the profile of the worker and **change your selection of information, press** button ←.

[next page]

HERE you can recheck the instructions (they will open in a new tab).

Now you will look through a worker's profile. Please, **select** the pieces of **information** that you would like to **provide to a manager** who will consider this worker for the **task** that consists of a **series of financial decisions**.

ОЛЕКС (OLEXIY) (ID 347)

SUMMARY (based on self-evaluation)

- ☐ Olexiy is even-tempered. He is good at solving difficult and complex problems and is creative.
- ☐ Olexiy is sometimes indecisive and fears mathematics.
- ☐ According to Olexiy, he could very probably convince others of his opinion in financial services.

DEMOGRAPHICS

- ☐ **Age:** 44
- ☐ **Marital status:** married
- ☐ **Number of children:** 1
- ☐ **Driving license:** yes
- ☐ **Number of completed online questionnaires:** 6 More information
  [After clicking on "More information":]
  The number of completed online questionnaires is a record from a database about the actual total number of online surveys that the worker has properly completed in

the past. Note: This button with more information will never be displayed to the manager.

EDUCATION

- ☐ **Level:** university – master's degree
- ☐ **Area of studies:** social geography
- ☐ **Favorite subject:** geography
- ☐ **Favorite subject:** English

WORK

- ☐ **Employment sector:** insurance
- ☐ **Current position:** product manager More information
  [After clicking on "More information":]
  Product manager is responsible for having an overview of the market, monitoring current trends and their identification. Based on the observations, he/she then creates strategic plans, including the design, creation and launch of new products. Note: This button with more information will never be shown to the manager.
- ☐ **Work experience in the last position:** 3 years
- ☐ **Job responsibilities:** product management
- ☐ **Job responsibilities:** content on intranet and web
- ☐ **Job responsibilities:** organization of testing of new products

CERTIFICATES (based on real tasks)

- ☐ Attained **2 points in a math test (average** of all candidates: **4** points) More information
  [After clicking on "More information":]
  The Math test included 10 questions that tested the knowledge of basic Math operations, equation solving, etc. Participants had 2.5 minutes to complete the test. Note: This button with more information will never be displayed to the manager.
- ☐ Attained **5 points in a financial-literacy quiz (average** of all candidates: **3.9** points) More information

[After clicking on "More information":]

The financial-literacy quiz included 5 questions that tested the understanding of basic financial concepts (inflation, interest, etc.). Note: This button with more information will never be displayed to the manager.

☐ Attained **41 points in a slider task (average** of all candidates: **24** points)
More information

[After clicking on "More information":]

The slider task is a mechanical task in which participants had to center within a 2-minute limit as many sliders as possible (max. 48) with a random initial position. Note: This button with more information will never be displayed to the manager.

VOLUNTEERING (based on real decisions)

☐ **Donated own participation fee** in **100%** of completed online surveys

SKILLS

☐ **Microsoft Word:** basic knowledge

☐ **Microsoft Excel:** basic knowledge

☐ **Microsoft PowerPoint:** basic knowledge

☐ **Internet banking:** using

☐ **English language:** good knowledge

☐ **Has experience with** product management

☐ **Has experience with** with holding stocks and mutual funds

INTERESTS

☐ Reading books

☐ Finance/business/economics

[next page]

Based on **your** earlier **selection**, the hiring **manager will see** the following information:
***

OJIEKC (OLEXIY) (ID 347)

[At this point, the assistant saw what would be displayed to the manager about this worker based on the assistant's choices.]

***

If you want to return to the profile of the worker and **change your selection of information, press** button ←.

[next page]

HERE you can recheck the instructions (they will open in a new tab).

Now you will look through a worker's profile. Please, **select** the pieces of **information** that you would like to **provide to a manager** who will consider this worker for the **task** that consists of a **series of financial decisions**.

PETR (ID 778)

SUMMARY (based on self-evaluation)

- ☐ Petr's strengths are logical thinking and trustworthiness.
- ☐ Petr is sometimes unorganized and postpones things.
- ☐ According to Petr, it is important to keep learning new things.

DEMOGRAPHICS

- ☐ **Age:** 38
- ☐ **Marital status:** married
- ☐ **Number of children:** 2
- ☐ **Driving license:** yes
- ☐ **Number of completed online questionnaires:** 7 More information
  [After clicking on "More information":]
  The number of completed online questionnaires is a record from a database about the actual total number of online surveys that the worker has properly completed in the past. Note: This button with more information will never be displayed to the

manager.

EDUCATION

- ☐ **Level:** secondary (with school leaving exam)
- ☐ **Area of studies:** business and service management
- ☐ **Favorite subject:** commodity expertise More information
  [After clicking on "More information":]
  It enables orientation in the main assortment groups in accordance with valid legislation and the requirements of business practice, it clarifies the issue of consumer properties, quality, evaluation of goods, defects of goods, labeling, and professional sale of goods. Note: This button with more information will never be shown to the manager.
- ☐ **Favorite subject:** mathematics

WORK

- ☐ **Employment sector:** trade – purchase and sale of goods
- ☐ **Current position:** cashier
- ☐ **Work experience in the current position:** 1 year
- ☐ **Job responsibilities:** communication
- ☐ **Job responsibilities:** service
- ☐ **Job responsibilities:** goods

CERTIFICATES (based on real tasks)

- ☐ Attained **5 points in a math test (average** of all candidates: **4** points) More information
  [After clicking on "More information":]
  The Math test included 10 questions that tested the knowledge of basic Math operations, equation solving, etc. Participants had 2.5 minutes to complete the test. Note: This button with more information will never be displayed to the manager.
- ☐ Attained **4 points in a financial-literacy quiz (average** of all candidates: **3.9** points) More information

[After clicking on "More information":]

The financial-literacy quiz included 5 questions that tested the understanding of basic financial concepts (inflation, interest, etc.). Note: This button with more information will never be displayed to the manager.

☐ Attained **9 points in a slider task (average** of all candidates: **24** points)
More information

[After clicking on "More information":]

The slider task is a mechanical task in which participants had to center within a 2-minute limit as many sliders as possible (max. 48) with a random initial position. Note: This button with more information will never be displayed to the manager.

SKILLS

☐ **Microsoft Word:** basic knowledge

☐ **Microsoft Excel:** basic knowledge

☐ **Microsoft PowerPoint:** no experience

☐ **Internet banking:** not using

☐ **English language:** partial knowledge

☐ **Has experience with** customer service

INTERESTS

☐ Watching TV

☐ Trips to the countryside

[next page]

Based on **your** earlier **selection**, the hiring **manager will see** the following information:
***

PETR (ID 778)

[At this point, the assistant saw what would be displayed to the manager about this worker based on the assistant's choices.]

***

If you want to return to the profile of the worker and **change your selection of information, press** button ←.

[next page]

HERE you can recheck the instructions (they will open in a new tab).

Now you will look through a worker's profile. Please, **select** the pieces of **information** that you would like to **provide to a manager** who will consider this worker for the **task** that consists of a **series of financial decisions**.

ZDEŇKA (ID 459)

SUMMARY (based on self-evaluation)

- ☐ Zdeňka has logical thinking. She is responsible and courteous.
- ☐ Zdeňka is sometimes indecisive and lacks self-confidence.
- ☐ Zdeňka considers herself good at money management.

DEMOGRAPHICS

- ☐ **Age:** 30
- ☐ **Marital status:** single
- ☐ **Number of children:** 0
- ☐ **Driving license:** yes
- ☐ **Number of completed online questionnaires:** 151 More information
  [After clicking on "More information":]
  The number of completed online questionnaires is a record from a database about the actual total number of online surveys that the worker has properly completed in the past. Note: This button with more information will never be displayed to the manager.

EDUCATION

- ☐ **Level:** university – master's degree
- ☐ **Area of studies:** statistics

☐ **Favorite subject:** statistics

☐ **Favorite subject:** demographics

☐ **Favorite subject:** accounting

WORK

☐ **Employment sector:** marketing/management/advertising/media

☐ **Current position:** project field manager More information
[After clicking on "More information":]
Project field manager is responsible for smooth and efficient day-to-day progress of a project. He/She tries to learn and fulfill needs of clients, set goals and timelines, determine a budget, manage the work group, and control the progress of the project in order to meet standards and regulations. He/She also makes interim reports and evaluations and suggests improvements of processes. Note: This button with more information will never be shown to the manager.

☐ **Job responsibilities:** communication

☐ **Job responsibilities:** database management

☐ **Job responsibilities:** work organization

CERTIFICATES (based on real tasks)

☐ Attained **8 points in a math test (average** of all candidates: **4** points) More information
[After clicking on "More information":]
The Math test included 10 questions that tested the knowledge of basic Math operations, equation solving, etc. Participants had 2.5 minutes to complete the test. Note: This button with more information will never be displayed to the manager.

☐ Attained **4 points in a financial-literacy quiz (average** of all candidates: **3.9** points) More information
[After clicking on "More information":]
The financial-literacy quiz included 5 questions that tested the understanding of basic financial concepts (inflation, interest, etc.). Note: This button with more information will never be displayed to the manager.

☐ Attained **48 points in a slider task (average** of all candidates: **24** points) More information

[After clicking on "More information":]

The slider task is a mechanical task in which participants had to center within a 2-minute limit as many sliders as possible (max. 48) with a random initial position. Note: This button with more information will never be displayed to the manager.

VOLUNTEERING (based on real decisions)

☐ Completed a survey for free in order to **donate the money to a charity** More information

[After clicking on "More information":]

In the questionnaire, we asked the worker whether he or she is willing to participate in another survey in upcoming days and donate the reward from participation to a charity of own choice. If the worker agreed, he or she received later an invitation for a survey in which it was explicitly mentioned that the reward will be donated. We could verify whether the worker truly completed this survey. Note: This button with more information will never be displayed to the manager.

☐ **Donated own participation fee** in **69%** of completed online surveys

SKILLS

☐ **Microsoft Word:** advanced

☐ **Microsoft Excel:** advanced

☐ **Microsoft PowerPoint:** basic knowledge

☐ **Internet banking:** using

☐ **English language:** good knowledge

☐ **Has experience with** mathematics

☐ **Has experience with** data entry

☐ **Has experience with** data analysis

INTERESTS

☐ Sport activities

☐ Music

[next page]

Based on **your** earlier **selection**, the hiring **manager will see** the following information:
***

ZDEŇKA (ID 459)

[At this point, the assistant saw what would be displayed to the manager about this worker based on the assistant's choices.]

***

If you want to return to the profile of the worker and **change your selection of information, press** button ←.

[next page]

HERE you can recheck the instructions (they will open in a new tab).

Now you will look through a worker's profile. Please, **select** the pieces of **information** that you would like to **provide to a manager** who will consider this worker for the **task** that consists of a **series of financial decisions**.

АНАТОЛИЙ (ANATOLIY) (ID 235)

SUMMARY (based on self-evaluation)

☐ Anatoliy has a technically-oriented mindset and is hungry for knowledge.

☐ Anatoliy is sometimes direct in expressing controversial opinions and unwilling to comply with social norms.

☐ Anatoliy considers himself good at money management and he does not leave financial decisions to other family members.

DEMOGRAPHICS

☐ **Age:** 38

☐ **Marital status:** single

☐ **Number of children:** 0

☐ **Driving license:** yes

☐ **Number of completed online questionnaires:** 58 More information

[After clicking on "More information":]

The number of completed online questionnaires is a record from a database about the actual total number of online surveys that the worker has properly completed in the past. Note: This button with more information will never be displayed to the manager.

EDUCATION

☐ **Level:** university – master's degree

☐ **Area of studies:** electronics and communication technology

☐ **Favorite subject:** telecommunication networks

☐ **Favorite subject:** circuit theory More information

[After clicking on "More information":]

An electrical circuit is a conductive connection of electrical elements, e.g. resistors, diodes, and switches. Circuit theory applies physical laws and principles in the analysis of elementary phenomena in DC and AC electrical circuits, defines basic circuit quantities (voltage, current) and basic circuit elements modeling all kinds of real energy interactions. The basic goal is the ability to calculate voltage and current anywhere in the circuit and based on them to assess the properties of electrical equipment. Note: This button with more information will never be displayed to the manager.

☐ **Favorite subject:** programming

WORK

☐ **Employment sector:** education

☐ **Current position:** IT administrator

☐ **Job responsibilities:** administration of computer network

☐ **Job responsibilities:** hardware maintenance

☐ **Work experience in the last position:** 5 years

CERTIFICATES (based on real tasks)

- ☐ Attained **3 points in a math test** (**average** of all candidates: **4** points) More information
  [After clicking on "More information":]
  The Math test included 10 questions that tested the knowledge of basic Math operations, equation solving, etc. Participants had 2.5 minutes to complete the test. Note: This button with more information will never be displayed to the manager.

- ☐ Attained **5 points in a financial-literacy quiz** (**average** of all candidates: **3.9** points) More information
  [After clicking on "More information":]
  The financial-literacy quiz included 5 questions that tested the understanding of basic financial concepts (inflation, interest, etc.). Note: This button with more information will never be displayed to the manager.

- ☐ Attained **12 points in a slider task** (**average** of all candidates: **24** points) More information
  [After clicking on "More information":]
  The slider task is a mechanical task in which participants had to center within a 2-minute limit as many sliders as possible (max. 48) with a random initial position. Note: This button with more information will never be displayed to the manager.

VOLUNTEERING (based on real decisions)

- ☐ **Donated own participation fee** in **16%** of completed online surveys

SKILLS

- ☐ **Microsoft Word:** professional
- ☐ **Microsoft Excel:** professional
- ☐ **Microsoft PowerPoint:** basic knowledge
- ☐ **Internet banking:** using
- ☐ **English language:** good knowledge
- ☐ **Has experience with** economics
- ☐ **Has experience with** mathematics

☐ **Has experience with** holding stocks and mutual funds

INTERESTS

☐ Reading books

☐ Gardening

☐ News about finance/business/economics

[next page]

Based on **your** earlier **selection**, the hiring **manager will see** the following information:
***

АНАТОЛИЙ (ANATOLIY) (ID 235)

[At this point, the assistant saw what would be displayed to the manager about this worker based on the assistant's choices.]

***

If you want to return to the profile of the worker and **change your selection of information, press** button ←.

[next page]

HERE you can recheck the instructions (they will open in a new tab).

Now you will look through a worker's profile. Please, **select** the pieces of **information** that you would like to **provide to a manager** who will consider this worker for the **task** that consists of a **series of financial decisions**.

ОЛЕСЯ (OLESYA) (ID 585)

SUMMARY (based on self-evaluation)

☐ Olesya's strengths are courtesy and flexibility.

☐ Olesya is sometimes direct in expressing controversial opinions and has bad performance under pressure.

☐ Olesya considers herself good at money management and certainly does not leave financial decisions to other family members.

☐ According to Olesya, people should try again when they do not succeed the first time.

DEMOGRAPHICS

☐ **Age:** 34

☐ **Marital status:** single

☐ **Number of children:** 0

☐ **Driving license:** yes

☐ **Number of completed online questionnaires:** 9 More information

[After clicking on "More information":]

The number of completed online questionnaires is a record from a database about the actual total number of online surveys that the worker has properly completed in the past. Note: This button with more information will never be displayed to the manager.

EDUCATION

☐ **Level:** secondary (with school leaving exam)

☐ **Area of studies:** trade

☐ **Favorite subject:** law

☐ **Favorite subject:** accounting

WORK

☐ **Employment sector:** advertising

☐ **Current position:** project manager More information

[After clicking on "More information":]

Project manager proposes a structure and staffing of the implementation team for a specific project. He/She is then in charge of this project, divides everything into sub-tasks, and then checks and supervises their fulfillment. While working on the project, he/she cooperates in determining the financial requirements of the project, makes time estimates and updates them. He/she regularly prepares written reports

on the status of the project. Note: This button with more information will never be shown to the manager.

☐ **Work experience in the current position:** 12 years

☐ **Job responsibilities:** communication with government offices

☐ **Job responsibilities:** invoicing

☐ **Job responsibilities:** communication with government

CERTIFICATES (based on real tasks)

☐ Attained **2 points in a math test** (**average** of all candidates: **4** points) More information

[After clicking on "More information":]

The Math test included 10 questions that tested the knowledge of basic Math operations, equation solving, etc. Participants had 2.5 minutes to complete the test. Note: This button with more information will never be displayed to the manager.

☐ Attained **5 points in a financial-literacy quiz** (**average** of all candidates: **3.9** points) More information

[After clicking on "More information":]

The financial-literacy quiz included 5 questions that tested the understanding of basic financial concepts (inflation, interest, etc.). Note: This button with more information will never be displayed to the manager.

☐ Attained **17 points in a slider task** (**average** of all candidates: **24** points) More information

[After clicking on "More information":]

The slider task is a mechanical task in which participants had to center within a 2-minute limit as many sliders as possible (max. 48) with a random initial position. Note: This button with more information will never be displayed to the manager.

SKILLS

☐ **Microsoft Word:** basic knowledge

☐ **Microsoft Excel:** basic knowledge

☐ **Microsoft PowerPoint:** basic knowledge

☐ **Internet banking:** not using

☐ **English language:** partial knowledge

☐ **Has experience with** data entry

☐ **Has experience with** customer service

INTERESTS

☐ Walks with the dog

☐ Sport activities

☐ Reading the newspaper

[next page]

Based on **your** earlier **selection**, the hiring **manager will see** the following information:
***

ОЛЕСЯ (OLESYA) (ID 585)

[At this point, the assistant saw what would be displayed to the manager about this worker based on the assistant's choices.]

***

If you want to return to the profile of the worker and **change your selection of information, press** button ←.

[next page]

If you are among 50 randomly chosen participants of this survey and you answer **correctly** the following **two questions**, you will earn **extra 200 points**.

In your opinion, what is the **country of origin of the last** worker whose profile you just saw?

◯ Czech Republic

◯ Post-Soviet country (e.g., Russia, Ukraine)

In your opinion, what is the **gender of the last** worker whose profile you just saw?

◯ Man

◯ Woman

[next page]

What guided your information selection for the hiring manager? **What did you try to achieve** with your information selection?

How much did you think about the hiring **manager** when selecting information about the workers for him/her?

Click on the slider to show the number that indicates its current position.

Not at all **Very much**

[next page]

Thank you for filling in the main part of our questionnaire. Now we would like to ask you to answer a couple of additional questions.

What is your **current employment situation**?

◯ Employed full-time

◯ Employed part-time

◯ Self-employed

◯ Unemployed but looking for a job

◯ Student, apprentice

◯ On maternal/parental leave / taking care of children

◯ Retired and not working

◯ In household

◯ Other

◯ I do not know / I do not want to answer

[next page]

Do you have **experience working in a hiring team**, e.g. Have you ever worked as a **human resource officer**?

◯ Yes
◯ No

[next page]

Please think about the **total net income of your household**. As net income, consider the **total amount that you have at your disposal**, **after taxes**—your income from work, state support, interest, etc.

To which **category** does the **net monthly income of your household** belong (total income of all members of the household together, without income of roommates)?

◯ No income
◯ Less than 15,000 Czech crowns
◯ 15,001-30,000 Czech crowns
◯ 30,001-40,000 Czech crowns
◯ 40,001-50,000 Czech crowns
◯ 50,001-75,000 Czech crowns
◯ 75,001-100,000 Czech crowns
◯ 100,001 and more Czech crowns
◯ I do not know / I do not want to answer

[new page]

Would you **mind** having as **your neighbor**:

To what extent do you **agree** with the following statements?

Foreigners from the countries of the former Soviet Union and Asia that are living long-term in the Czech Republic...

	Definitely would mind	Somewhat would mind	Indifferent	Rather would NOT mind	Definitely would NOT mind
Czech	◯	◯	◯	◯	◯
Russian	◯	◯	◯	◯	◯
Ukrainian	◯	◯	◯	◯	◯
Chinese	◯	◯	◯	◯	◯
Mongol	◯	◯	◯	◯	◯
Indian	◯	◯	◯	◯	◯

	Totally agree	Agree	I do not have an opinion	DISagree	Totally DIS-agree
present health risks (spreading diseases)	◯	◯	◯	◯	◯
cause criminality to increase	◯	◯	◯	◯	◯
threaten our way of life	◯	◯	◯	◯	◯
increase total unemployment	◯	◯	◯	◯	◯

To what extent do you **agree** with the following statements?

Foreigners **from the countries of the former Soviet Union and Asia** that are living long-term in the Czech Republic...

	Totally DIS-agree	DISgree	I do not have an opinion	Agree	Totally agree
help in resolving the problem of the ageing population	◯	◯	◯	◯	◯
contribute to developing the economy	◯	◯	◯	◯	◯
enrich our own culture	◯	◯	◯	◯	◯

[new page]

To what extent do you **agree** with the following statements?

	Totally agree	Agree	I do not have an opinion	DISagree	Totally DISagree
Women should always prioritize family over career.	◯	◯	◯	◯	◯
Women should take maternal leave after childbirth, not men.	◯	◯	◯	◯	◯
Women should take care of the household more than men.	◯	◯	◯	◯	◯
Women should take care of children more than men.	◯	◯	◯	◯	◯

To what extent do you **agree** with the following statements?

	Totally agree	Agree	I do not have an opinion	DISagree	Totally DISagree
Men are better managers than women.	◯	◯	◯	◯	◯
Financial provision for the family is foremost men's concern.	◯	◯	◯	◯	◯
Boys are more talented in technical fields and maths than girls.	◯	◯	◯	◯	◯

[next page]

Thank you for your participation. If you have any comments or questions concerning this survey, please write them in the field below. Your feedback is very important to us so that we can keep improving our research.

Bibliography

Almås, Ingvild, Alexander W Cappelen, and Bertil Tungodden. 2020. "Cutthroat capitalism versus cuddly socialism: Are Americans more meritocratic and efficiency-seeking than Scandinavians?" *Journal of Political Economy* 128 (5): 1753–1788.

Alonso, Ricardo, and Odilon Camara. 2016. "Bayesian persuasion with heterogeneous priors." *Journal of Economic Theory* 165:672–706.

Arieli, Itai, Yakov Babichenko, Rann Smorodinsky, and Takuro Yamashita. 2020. "Optimal Persuasion via Bi-Pooling." *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20. New York, NY, USA: Association for Computing Machinery, 641.

Artiga González, Tanja, Francesco Capozza, and Georg D Granic. 2022. "Political Support, Cognitive Dissonance and Political Preferences." Cesifo working paper No. 9549.

Baron, Jonathan. 2012. "The point of normative models in judgment and decision making." *Frontiers in Psychology* 3:577.

Barron, Kai, Ruth Ditlmann, Stefan Gehrig, and Sebastian Schweighofer-Kodritsch. 2022. "Explicit and implicit belief-based gender discrimination: A hiring experiment." Cesifo working paper No. 9731.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka. 2016. "Attention discrimination: Theory and field experiments with monitoring information acquisition." *American Economic Review* 106 (6): 1437–75.

Becker, Sascha O, Ana Fernandes, and Doris Weichselbaumer. 2019. "Discrimination in hiring based on potential and realized fertility: Evidence from a large-scale field experiment." *Labour Economics* 59:139–152.

Bertogg, Ariane, Christian Imdorf, Christer Hyggen, Dimitris Parsanoglou, and Rumiana Stoilova. 2020. "Gender discrimination in the hiring of skilled professionals in two male-dominated occupational fields: a factorial survey experiment with real-world vacancies and recruiters in four European countries." *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 72 (1): 261–289.

Bertrand, Marianne, and Esther Duflo. 2017. "Field experiments on discrimination." *Handbook of Economic Field Experiments* 1:309–393.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American Economic Review* 94 (4): 991–1013.

Bohren, J Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope. 2019. "Inaccurate statistical discrimination: An identification problem." Nber working paper No. 25935.

Bohren, J Aislinn, Peter Hull, and Alex Imas. 2022. "Systemic discrimination: Theory and measurement." Nber working paper No. 29820.

Bohren, J Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The dynamics of discrimination: Theory and evidence." *American Economic Review* 109 (10): 3395–3436.

Boldt, Annika, Charles Blundell, and Benedetto De Martino. 2019. "Confidence modulates exploration and exploitation in value-based learning." *Neuroscience of consciousness* 2019 (1): niz004.

Bradler, Christiane, Susanne Neckermann, and Arne Jonas Warnke. 2019. "Incentivizing creativity: A large-scale experiment with performance bonuses and gifts." *Journal of Labor Economics* 37 (3): 793–851.

Brewer, Neil, and Anne Burke. 2002. "Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments." *Law and Human Behavior* 26 (3): 353–364.

Brock, J Michelle, and Ralph De Haas. forthcoming. "Discriminatory Lending: Evidence from Bankers in the Lab." *American Economic Journal: Applied Economics*.

Cappelen, Alexander W, Ranveig Falch, and Bertil Tungodden. 2019. "The boy crisis: Experimental evidence on the acceptance of males falling behind." Nhh dept. of economics discussion paper No. 06/2019.

Cappelen, Alexander W, Sebastian Fest, Erik Ø Sørensen, and Bertil Tungodden. 2020. "Choice and personal responsibility: What is a morally relevant choice?" *Review of Economics and Statistics*, pp. 1–35.

Chernoff, Herman. 1961. "Sequential tests for the mean of a normal distribution." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1. University of California Press, 79–91.

Crawford, Vincent P, and Joel Sobel. 1982. "Strategic information transmission." *Econometrica: Journal of the Econometric Society*, pp. 1431–1451.

Cunningham, Tom, and Jonathan de Quidt. 2022. "Implicit preferences." Cepr discussion paper DP17343.

Czech Statistical Office. 2021. Míry zaměstnanosti, nezaměstnanosti a ekonomické aktivity - prosinec 2021 [Rates of employment, unemployment, and economic activity - December 2021]. https://www.czso.cz/csu/czso/household-income-and-living-conditions-6yp06pfzwa.

Desender, Kobe, Annika Boldt, and Nick Yeung. 2018. "Subjective confidence predicts information seeking in decision making." *Psychological science* 29 (5): 761–778.

Dobbin, Frank, and Alexandra Kalev. 2016. "Why diversity programs fail." *Harvard Business Review* 94 (7): 14.

Dupas, Pascaline, Alicia Sasser Modestino, Muriel Niederle, Justin Wolfers, et al. 2021. "Gender and the dynamics of economics seminars." Nber working paper No. 28494.

Dworczak, Piotr, and Giorgio Martini. 2019. "The simple economics of optimal persuasion." *Journal of Political Economy* 127 (5): 1993–2048.

Dye, Ronald A. 1985. "Disclosure of nonproprietary information." *Journal of Accounting Research*, pp. 123–145.

Eberhardt, Markus, Giovanni Facchini, and Valeria Rueda. 2022. "Gender Differences in Reference Letters: Evidence from the Economics Job Market." Cepr discussion paper No. DP16960.

Enke, Benjamin, and Thomas Graeber. 2022. "Cognitive uncertainty." Unpublished manuscript.

Enke, Benjamin, Thomas Graeber, and Ryan Oprea. 2022. "Confidence, Self-Selection, and Bias in the Aggregate." Unpublished manuscript.

Esponda, Ignacio, Ryan Oprea, and Sevgi Yuksel. 2023. "Seeing What is Representative." *Quarterly Journal of Economics*, 05, qjad020.

Fleming, Stephen M, and Nathaniel D Daw. 2017. "Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation." *Psychological Review* 124 (1): 91–114.

Folke, Tomas, Catrine Jacobsen, Stephen M Fleming, and Benedetto De Martino. 2016. "Explicit representation of confidence informs future value-based decisions." *Nature Human Behaviour* 1 (1): 1–8.

Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki. 2018. "Speed, accuracy, and the optimal timing of choices." *American Economic Review* 108 (12): 3651–84.

Gallen, Yana, and Melanie Wasserman. 2021. "Informed Choices: Gender Gaps in Career Advice." Cepr discussion paper No. DP15728.

Galperti, Simone. 2019. "Persuasion: The art of changing worldviews." *American Economic Review* 109 (3): 996–1031.

Gentzkow, Matthew, and Emir Kamenica. 2016. "A Rothschild-Stiglitz approach to Bayesian persuasion." *American Economic Review* 106 (5): 597–601.

Gill, David, and Victoria Prowse. 2012. "A structural analysis of disappointment aversion in a real effort competition." *American Economic Review* 102 (1): 469–503.

———. 2019. "Measuring costly effort using the slider task." *Journal of Behavioral and Experimental Finance* 21:1–9.

Glassdoor. 2021. Hiring Managers vs. Recruiters: What's the Difference? https://www.glassdoor.com/employers/blog/hiring-manager-vs-recruiter-relationships-mind-the-gap/.

González, M José, Clara Cortina, and Jorge Rodríguez. 2019. "The role of gender stereotypes in hiring: a field experiment." *European Sociological Review* 35 (2): 187–204.

Hagenbach, Jeanne, Frédéric Koessler, and Eduardo Perez-Richet. 2014. "Certifiable Pre-Play Communication: Full Disclosure." *Econometrica* 82 (3): 1093–1131.

He, Haoran, Sherry Xin Li, and Yuling Han. forthcoming. "Labor Market Discrimination against Family Responsibilities: A Correspondence Study with Policy Change in China." *Journal of Labor Economics*.

Hengel, Erin. 2022. "Publishing While Female: Are Women Held to Higher Standards? Evidence from Peer Review." *Economic Journal*.

Hummel, Patrick, John Morgan, and Phillip C Stocken. 2018. "A model of voluntary managerial disclosure." Unpublished manuscript.

International Labor Organization. 2020. These occupations are dominated by women. https://ilostat.ilo.org/these-occupations-are-dominated-by-women/.
Kaas, Leo, and Christian Manger. 2012. "Ethnic discrimination in Germany's labour market: A field experiment." *German Economic Review* 13 (1): 1–20.
Kamenica, Emir. 2019. "Bayesian persuasion and information design." *Annual Review of Economics* 11:249–272.
Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian persuasion." *American Economic Review* 101 (6): 2590–2615.
Kessler, Judd B, Corinne Low, and Colin D Sullivan. 2019. "Incentivized resume rating: Eliciting employer preferences without deception." *American Economic Review* 109 (11): 3713–44.
Kline, Patrick, Evan K Rose, and Christopher R Walters. 2022. "Systemic discrimination among large US employers." *Quarterly Journal of Economics* 137 (4): 1963–2036.
Kolotilin, Anton. 2018. "Optimal information disclosure: A linear programming approach." *Theoretical Economics* 13 (2): 607–635.
Kolotilin, Anton, and Alexander Wolitzky. 2020. "Assortative Information Disclosure." UNSW Economics Working Paper 2020-08.
Krajbich, Ian, Bastiaan Oud, and Ernst Fehr. 2014. "Benefits of neuroeconomic modeling: New policy interventions and predictors of preference." *American Economic Review* 104 (5): 501–06.
Kübler, Dorothea, Julia Schmid, and Robert Stüber. 2018. "Gender discrimination in hiring across occupations: a nationally-representative vignette study." *Labour Economics* 55:215–229.
Le Gall, Jean-François. 2016. *Brownian motion, martingales, and stochastic calculus*. Springer.
Liptser, RS, and AN Shiryaev. 2000. *Statistics of Random Processes, 2nd edn. I and II*. Springer, Berlin.
Lusardi, Annamaria, Olivia S Mitchell, and Vilsa Curto. 2010. "Financial literacy among the young." *Journal of Consumer Affairs* 44 (2): 358–380.
Marr, D. 1982. *Vision*. New York, NY: WH Freeman.
Milgrom, Paul, and John Roberts. 1986. "Relying on the information of interested parties." *The RAND Journal of Economics*, pp. 18–32.
Milgrom, Paul R. 1981. "Good news and bad news: Representation theorems and applications." *The Bell Journal of Economics*, pp. 380–391.
Miura, Shintaro. 2018. "Prudence in Persuasion." Unpublished manuscript.
Moran, Rani, Andrei R Teodorescu, and Marius Usher. 2015. "Post choice information integration as a causal determinant of confidence: Novel data and a computational account." *Cognitive psychology* 78:99–147.
Morris, Stephen, and Philipp Strack. 2019. "The Wald problem and the relation of sequential sampling and ex-ante information costs." Working paper.
Oesch, Daniel. 2020. "Discrimination in the hiring of older jobseekers: Combining a survey experiment with a natural experiment in Switzerland." *Research in Social Stratification and Mobility* 65:100441.
Øksendal, Bernt. 2003. *Stochastic differential equations: an introduction with applications*. Springer.
Peskir, Goran, and Albert Shiryaev. 2006. *Optimal stopping and free-boundary problems*.

Springer.

Petit, Pascale. 2007. "The effects of age and family constraints on gender hiring discrimination: A field experiment in the French financial sector." *Labour Economics* 14 (3): 371–391.

Phelps, Edmund S. 1972. "The statistical theory of racism and sexism." *American Economic Review* 62 (4): 659–661.

Pleskac, Timothy J, and Jerome R Busemeyer. 2010. "Two-stage dynamic signal detection: a theory of choice, decision time, and confidence." *Psychological Review* 117 (3): 864–901.

Pouget, Alexandre, Jan Drugowitsch, and Adam Kepecs. 2016. "Confidence and certainty: distinct probabilistic quantities for different goals." *Nature neuroscience* 19 (3): 366–374.

Purcell, Braden A, and Roozbeh Kiani. 2016. "Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy." *Proceedings of the National Academy of Sciences* 113 (31): E4531–E4540.

Quillian, Lincoln, John J Lee, and Mariana Oliver. 2020. "Evidence from field experiments in hiring shows substantial additional racial discrimination after the callback." *Social Forces* 99 (2): 732–759.

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H Midtbøen. 2017. "Meta-analysis of field experiments shows no change in racial discrimination in hiring over time." *Proceedings of the National Academy of Sciences* 114 (41): 10870–10875.

Rahnev, Dobromir, Tarryn Balsdon, Lucie Charles, Vincent De Gardelle, Rachel Denison, Kobe Desender, Nathan Faivre, Elisa Filevich, Stephen Fleming, Janneke Jehee, et al. 2021. "Consensus goals for the field of visual metacognition." https://doi.org/10.31234/osf.io/z8v5x.

Rahnev, Dobromir, Kobe Desender, Alan LF Lee, William T Adler, David Aguilar-Lleyda, Başak Akdoğan, Polina Arbuzova, Lauren Y Atlas, Fuat Balcı, Ji Won Bang, et al. 2020. "The confidence database." *Nature human behaviour* 4 (3): 317–325.

Sah, Sunita, Don A Moore, and Robert J MacCoun. 2013. "Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness." *Organizational Behavior and Human Decision Processes* 121 (2): 246–255.

Schulz, Lion, Stephen M Fleming, and Peter Dayan. 2021. "Metacognitive computations for information search: Confidence in control." *bioRxiv*.

Seidmann, Daniel J, and Eyal Winter. 1997. "Strategic information transmission with verifiable messages." *Econometrica: Journal of the Econometric Society*, pp. 163–169.

Shea, Nicholas, Annika Boldt, Dan Bang, Nick Yeung, Cecilia Heyes, and Chris D Frith. 2014. "Supra-personal cognitive control and metacognition." *Trends in cognitive sciences* 18 (4): 186–193.

Shin, Hyun Song. 1994. "The burden of proof in a game of persuasion." *Journal of Economic Theory* 64 (1): 253–264.

Shiryaev, Albert N. 2007. *Optimal rules*. 2nd ed. Translated by A.B. Aries. New York: Springer. Originally published as *Optimal'nye pravila ostanovski.* (Moscow: Nauka, 1969).

Tajima, Satohiro, Jan Drugowitsch, and Alexandre Pouget. 2016. "Optimal policy for value-based decision-making." *Nature communications* 7 (1): 1–12.

Van Borm, Hannah, and Stijn Baert. 2022. "Diving in the minds of recruiters: What triggers gender stereotypes in hiring?" Iza discussion paper No. 15261.

Van den Berg, Ronald, Ariel Zylberberg, Roozbeh Kiani, Michael N Shadlen, and Daniel M Wolpert. 2016. "Confidence is the bridge between multi-stage decisions." *Current Biology* 26 (23): 3157–3168.

Vullioud, Colin, Fabrice Clément, Thom Scott-Phillips, and Hugo Mercier. 2017. "Confidence as an expression of commitment: Why misplaced expressions of confidence backfire." *Evolution and Human Behavior* 38 (1): 9–17.

Wald, Abraham. 1947. *Sequential analysis*. New York: John Wiley & Sons.

Wu, Alice H. 2018. "Gendered language on the economics job market rumors forum." *AEA Papers and Proceedings*, Volume 108. 175–79.

Yeung, Nick, and Christopher Summerfield. 2012. "Metacognition in human decision-making: confidence and error monitoring." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1594): 1310–1321.

Yin, Dezhi, Sabyasachi Mitra, and Han Zhang. 2016. "Research note—When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth." *Information Systems Research* 27 (1): 131–144.

Zhitlukhin, M V, and A A Muravlev. 2013. "On Chernoff's Hypotheses Testing Problem for the Drift of a Brownian Motion." *Theory of Probability & Its Applications* 57 (4): 708–717.