The London School of Economics and Political Science

Essays on Misspecified Models

Heidi Christina Thysen

A thesis submitted to the Department of Economics
for the degree of Doctor of Philosophy

30 April, 2021

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of approximately 29,000 words.

Statement of conjoint work and inclusion of previous work

I confirm that Chapter 1 was jointly co-authored with Heiner Schumacher, and I contributed 50% of this work. Proposition 7 of this Chapter was the result of previous study for a masters degree I undertook at Aarhus University.

I confirm that Chapter 2 was jointly co-authored with Kfir Eliaz and Ran Spiegler, and I contributed 33% of this work. A version of this Chapter is published in Journal of Economic Theory (2021), vol. 192, paper-id 105192.

Statement of use of third party for editorial help

I confirm that Chapter 3 was copy edited for conventions of language, spelling and grammar by Hugh Burrows.

Acknowledgement

My journey to do my PhD at LSE actually started two years prior to my enrolment, when I took a summer course taught by Kfir Eliaz. This put me on a new and exciting path and it would not have been possible without the patience and support of a lot of people.

Thanks to all of my supervisors, both the official and unofficial ones. Kfir, you took me under your wings, helped and supported me when I was first dabbling with research, and you have been there for me ever since. For that I am very grateful. Rani Spiegler, you have helped and supported me in every way possible. Officially, you just became my supervisor at the end of my PhD, but unofficially you have been there through it all. Without you I would not have been here and I appreciate all of our candid discussions more than you can possibly know. Gilat Levy, you have been the best supervisor I could have wished for. You are always ready with advice - both academically and otherwise. You have taught me so much - Thank you!

Many thanks to the faculty and students at Eitan Berglas School of Economics, who instantly made me feel at home there. In particular I owe a special thanks to Yair Antler, Benjamin Bachi, Ran Eilat and Alex Frug, and to the people outside the department that made my visits to Tel Aviv unforgettable such at Daniella, Lorenzo, Eleonora, Flavio and the rest of the Italian/Physics jellyfish. A big thanks goes to Debraj Ray for inviting me to NYU and NRET. I really enjoyed my semester there and you made me feel welcome from the first day. A warm thanks to all of the faculty and students at NYU for the many great discussions and experiences.

I am very grateful for all the faculty and students in the Micro Theory and the Political Science and Political Economics groups at LSE. A special thanks to Ronny Razin, Balazs Scentes, Stephane Wolton and Arduino Tomasi. Another big thanks goes to all the students that made my time at LSE and outside more enjoyable. In particular, Daniel Albuquerque, Jamie Coen, Alexandre Desbuquois, Andres Ramirez, Edoardo Leonardi, William Matcham, Akash Raja, Veronica Restrepo, Bilal Tabti, Cecilia Wood and Celine Zipfel. A special thanks to Lu Liu for all the discussions, lunches and work visits to LSE.

Thanks to the best co-authors I could have wished for: Kfir Eliaz, Andrew Ellis, Clement Minaudier, Heiner Schumacher and Rani Spiegler. You made the work enjoyable.

Last but not least a special thanks to my family and friends, who helped my through this whole process close and from afar. In particular my parents, my sisters, my grandparents, Wednesday Club, Racletten, Lea, Louise, Nina, Martin, Mille & Torben, Peter, Tanja and of course Seb. Seb, thanks for all the love and support and the many adventures to come.

Abstract

These essays examine how relaxing common assumptions affect the strategic interactions between agents. It investigates how the presence of an agent with a simplified causal model influences the contract and disclosure of information with other agents, as well as the impact of changes to the institutional settings. By theoretically modelling these choices, it aims to improve the understanding of equilibrium effects and thereby contributing to debates about the optimal design of contracts, strategic information transmission and political budget cycles and the impact of assumptions.

The first chapter analyses the contract between a firm owner and a employee, when the firm cannot observe the employee's action and the employee's belief about how her action influences the contractible variable is governed by a misspecified causal model. It contributes to the existing literature by explicitly modelling the source of the employee's misspecified beliefs. This approach allows us to shed light on the variables the firm owner would want to include in the contract given the employee's mistakes as well as the intermediate variables the employee needs to include in her causal model in order to act as if she understands how her action influence the contractible variable.

The second chapter examines how an informed agent conveys information to an uninformed agent when he can simultaneously influence the messages she receives and how she interprets them. This relaxes the assumption that agents always understand the meaning of messages in equilibrium.

The third chapter analyses how political budget cycles change when the politician in charge can choose to call for a snap election in periods before the end of the term. This contributes to the existing literature by taking the equilibrium effects of early election into consideration and thereby the effect of the continuation value of being in office.

Contents

List of Figures

Chapter 1

Equilibrium Contracts and Boundedly Rational Expectations

1.1 Introduction

The canonical principal-agent model of contracting under asymmetric information assumes that the agent knows the probabilistic consequences of all available actions. Formally, these are defined by a production function $p(y \mid a)$, where y is the contractible output and a the agent's action. Given the incentives provided by the contract, the agent chooses an action that – according to this function – maximizes her expected payoff. However, in an organization, $p(y \mid a)$ is typically a complex object. It may reflect knowledge that is unavailable to the agent or that the agent cannot process due to cognitive limitations. Herbert Simon therefore proposed that administrative behavior may be "boundedly rational" (Simon, Simon, 1947, 1955).

The common approach to contracting with boundedly rational agents is to assume directly that beliefs $\hat{p}(y \mid a)$ about the production function are biased so that $\hat{p}(y \mid a) \neq p(y \mid a)$. This captures, for example, an agent's overconfidence. An important implication of this approach is that the optimal contract may exploit the agent, in the sense that her (true) expected payoff falls below her reservation utility (e.g., Kőszegi, 2014). However, it is unclear how sustainable biased beliefs – and hence exploitation – would be when the agent gathers experience.

In this paper, we apply a new approach where the agent derives her beliefs about $p(y \mid a)$ from the data generated by the true production process, the implemented strategy q, and

a non-parametric subjective model $\mathcal{R}$. A strategy q is a probability distribution over the agent's actions and a model $\mathcal{R}$ is a collection of variables and causal relationships between these variables. It captures what the agent knows about the production process. This model may be misspecified. For example, it may be "too simple" relative to the complexity of the organization: Empirical regularities that matter for the principal's project may not appear in $\mathcal{R}$. We derive the agent's subjective beliefs about $p(y \mid a)$ using Spiegler's (2016) Bayesian network framework; we denote them by $p_{\mathcal{R}}(y \mid a; q)$. An equilibrium contract implements a strategy q if it is optimal for the agent to follow q under this contract given her beliefs $p_{\mathcal{R}}(y \mid a; q)$. We study the properties of the optimal equilibrium contract, and obtain several new results on optimal contracting and organization.

Our framework captures a variety of misconceptions that even experienced decision makers may exhibit. Consider the basic management practice of inventory control. Its implementation reduces the working time spent on dealing with inputs that are not needed, which in turn increases productivity. However, if the manager does not have the causal chain "inventory control $\rightarrow$ working time allocation $\rightarrow$ productivity" on her mind, she may see no benefit from implementing inventory control, and choose a suboptimal organization of the workplace. Indeed, Bloom et al. (2013) document that the managers in several large Indian textile factories did not acknowledge the positive impact of basic management practices (like inventory control) on productivity. They only changed their mind after substantial consulting and after these measures proved effective.[1]

Another example is the choice of management style. Individuals who are appointed to a management position often struggle to find the right approach. Suppose a mid-level manager has to choose whether she closely controls her subordinates' actions ("micromanagement"). This reduces misbehavior, but it also diminishes her subordinates' performance (DeCaro et al., 2011). Nevertheless, in the fog of business, the manager may only focus on reducing misbehavior and neglect employee motivation. Micromanagement then appears to her as more appealing than it really is, and she therefore may adopt an inefficient management style.

Finally, decision makers may not fully understand their clients. Consider a marketer whose job is to increase sales. One strategy to increase sales is to make cold-calls, that is, calling potential customers without prior consent. Making cold-calls improves consumers' information about the firm's product, but also reduces the firm's reputation since some

[1]There are a number of further well-documented cases where experienced decision makers ignore important aspects of their operation; see, e.g., Nuland (2004) or Hanna et al. (2014).

customers start doubting the quality of the product if such a marketing strategy is applied.[2] Sales increase both in consumer information and reputation. However, when choosing her action, the marketer may not take the firm's reputation into account. Then the only mechanism on her mind is that making cold-calls improves consumer information, and that more information translates into more sales. In all these examples, the decision makers arguably know the expected outcomes from their usual actions. They just may incorrectly infer the counter-factual consequences of a change in their behavior. This is what we can capture in our framework.

The Bayesian network approach roughly works as follows[3] in the marketer example (which we use as running example throughout the paper). The setting describes an "extended production function" $p(x_1, x_2, y \mid a)$, i.e., a joint probability distribution over the realization of consumer information x_1, reputation x_2, and sales y for any given action a. This function reflects the objective model $\mathcal{R}^*$ of the project: $\mathcal{R}^*$ contains all relevant variables, {action, consumer information, reputation, sales}, and the causal relationships between these variables. The agent's subjective model $\mathcal{R}$ is a simplified version of $\mathcal{R}^*$ as it only contains the variables {action, consumer information, sales}, and their causal relationships. Her beliefs are derived by fitting $\mathcal{R}$ to the objective probability distribution, which is generated by the implemented strategy q and the extended production function $p(x_1, x_2, y \mid a)$. Thus, the different elements in the agent's subjective model $\mathcal{R}$ are quantified using input from the true data-generating process. Combining these elements yields the agent's subjective beliefs $p_{\mathcal{R}}(y \mid a; q)$, which in general are not invariant to changes in q.

We show that the optimal equilibrium contract exhibits the following features. First, a weak restriction on the agent's subjective model guarantees that the participation constraint is not affected. This restriction is that $\mathcal{R}$ is "perfect", which means that the agent takes into account the link between any two variables in $\mathcal{R}$ that have a joint influence on a third variable in $\mathcal{R}$. She then correctly predicts the marginal equilibrium distribution over output (Spiegler, 2017), so that the optimal equilibrium contract does not exploit the agent. Importantly, a perfect $\mathcal{R}$ ensures in many cases that there are no informational cues in the data the agent gathers on the equilibrium path that could alert her about the misspecification in $\mathcal{R}$.

Second, the principal may strictly benefit from the misspecification in the agent's model even when exploitation is infeasible. In the marketer example, if the principal implements

[2] This mechanism is called "demarketing" (Miklós-Thal and Zhang, 2013): Extensive marketing can backfire since it may be interpreted as a signal for low quality.

[3] Missing technical details will be explained thoroughly in the next section.

making cold-calls, then, by not taking reputation into account, the agent overestimates the drop in sales after deviation to not making cold-calls, i.e., she is "control optimistic" as defined by Spinnewijn (2013). This relaxes the incentive compatibility constraint, so that the principal can implement cold-calls with fewer incentives than if the agent had rational expectations.

Third, when $\mathcal{R}$ is perfect, the incentive scheme in the optimal equilibrium contract appears to the agent as optimal for the principal. The agent then cannot deduct from the shape of incentives that her beliefs are biased. This is again different from the optimal contract under exogenously given biased beliefs where the agent may notice that the principal is betting against her. We show that in some cases the optimal equilibrium contract is "justifiable", i.e., it is optimal for the principal from the agent's point of view.

Taken together, these results show that an agent's misperceptions can be sustainable in an organizational context: Neither her experiences on the equilibrium path nor the shape of the incentive contract inform the agent about the mistake in her thinking, and the principal benefits from this mistake. Building on these insights, we further analyze three topics in organizational economics: First, we derive a behavioral version of the informativeness principle. Second, we characterize when misspecifications in the agent's model affect her beliefs. And third, we revisit the trade-off between risk and incentives. We briefly describe each topic in turn.

An important question in contract theory is on which variables the optimal contract should condition the agent's wage. According to the informativeness principle (e.g., Holmström, 1979, Chaigneau et al., 2019), the optimal contract conditions on an additional signal z only if z provides information about the agent's action that is not contained in y. We can derive an analogous statement when the agent has correct expectations on the equilibrium path about the joint distribution of y and z (with a further qualification this holds if $\mathcal{R}$ is perfect). In this case, the optimal equilibrium contract conditions on z only if the agent's action a and z are not independent conditional on y according to the agent's subjective beliefs. This result does not depend on other properties of the agent's subjective model $\mathcal{R}$, and hence would hold in any setting where the agent's beliefs about the joint distribution of y and z are correct. Nevertheless, we can use results from the Bayesian network literature to state sufficient conditions on $\mathcal{R}$ so that the result's requirements are satisfied. We apply these findings to provide a new explanation for why executive compensation contracts often do not condition on peer-performance (e.g., Bertrand and Mullainathan, 2001, Bebchuk and Fried, 2004).

Next, misspecifications in $\mathcal{R}$ do not always affect the agent's beliefs and optimal equilibrium contract. The agent is "behaviorally rational" if she correctly anticipates the production function, or, formally, $p_{\mathcal{R}}(y \mid a; q) = p(y \mid a)$ for all possible a and q, regardless of the parametrization of the extended production function. We can find a correspondence $H^*(\mathcal{R}^*)$ which indicates for a given objective model $\mathcal{R}^*$ the set of variables the agent must take into account in her simplified subjective model $\mathcal{R}$ so that she is behaviorally rational. We show that $H^*(\mathcal{R}^*)$ is often a strict subset of the variables in $\mathcal{R}^*$, and that the difference between a variable $i \in H^*(\mathcal{R}^*)$ and a variable $j \not\in H^*(\mathcal{R}^*)$ can be quite nuanced.

The characterization of $H^*(\mathcal{R}^*)$ shows which variables matter for the agent's beliefs. An important interpretation of the objective model $\mathcal{R}^*$ is that it captures the agent's job, i.e., through which tasks, interactions, and decision-making powers she influences the final output. We can have two extended production functions that give rise to the same "reduced-form" production function $p(y \mid a)$, but that differ in their causal model $\mathcal{R}^*$, and hence in the extent to which simplifications affect $p_{\mathcal{R}}(y \mid a; q)$. This allows us to examine which organizational features potentially cause the agent to overestimate the productivity of her effort. Consider an agent in a management position in which her effort influences the behavior of other workers (e.g., a group of marketers). If the agent does not understand the difficulties of their job (e.g., that cold-calls have a partial negative effect on sales through their effect on firm reputation), she overestimates her subordinates' – and hence her own – productivity. There are different instances where this could happen: The agent may be a technical expert who is promoted into a management position in which she oversees the actions of workers whose job she does not fully understand. Alternatively, it may be the case that subordinates do not communicate the problems they face to their managers (due to career concerns). These phenomena are usually discussed critically in the management literature (e.g., Porter et al., 2004), but in our framework they advance the agent's effort motivation and hence benefit the principal.

Finally, our framework allows for comparative statics since the agent's beliefs are derived from the parameters of the true production process. We briefly revisit the trade-off between risk and incentives, which has been extensively debated both in the theoretical and empirical contract theory literature (e.g., Prendergast, 2002). We show that when the agents subjective model is misspecified, then there can be a positive association between risk and the level of incentives the optimal equilibrium contract provides.

Related Literature. Our basic model is the principal-agent framework introduced by Holmström (1979) and Grossman and Hart (1983). Holmström (1979) states a version of the informativeness principle. A generalization of it can be found in, e.g., Chaigneau et al. (2019). In the canonical framework, both principal and agent know the production function $p(y \mid a)$.

There are different approaches in behavioral contract theory that relax the assumption of unbiased beliefs about $p(y \mid a)$. First, several contracting models directly assume that the agent's beliefs about the production function are biased, i.e., $\hat{p}(y \mid a) \neq p(y \mid a)$; see Fang and Mocarini (2005), Van den Steen (2005), Gervais and Goldstein (2007), Santo-Pinto (2008), De la Rosa (2011), Sautmann (2007), Sautmann (2013), Spinnewijn (2013), Spinnewijn (2015). Specifically, this approach is used to model an overconfident agent who overestimates the probability of good states and underestimates the probability of bad states. This typically allows the principal to exploit the agent by paying more after high output and much less after low output, in which case the agent's expected payoff is below her reservation utility.

Second, a rich literature builds state-space models of "unawareness" (e.g., Dekel et al., 1998, Heifetz et al., 2006, Heifetz et al., 2013) and applies them to contracting settings. Auster (2013) examines a principal-agent model with an agent who is unaware of some output levels y, which again implies that the contract is exploitative. von Thadden and Zhao (2012) and von Thadden and Zhao (2014) assume that the agent is unaware of her available actions a and chooses a default action unless the principal educates her. Unawareness then relaxes incentive compatibility at the default action.

Third, in order to justify biased beliefs, several papers assume that the agent knows the link between action and outcomes $p(y \mid a)$, but potentially gains from holding biased beliefs. She then chooses beliefs $\hat{p}(y \mid a)$ that solve the trade-off between the losses from biased decision-making and the gains from managing a self-control problem (Bénabou and Tirole, 2002) or from enjoying anticipatory utility (Brunnermeier and Parker, 2005, Kőszegi, 2006). For an organizational context, Bénabou (2013) shows how the interaction between group members can make the suppression of bad news a strategic complement, so that collective denial of adverse signals ("groupthink") occurs in equilibrium. Immordino et al. (2015) show that if anticipatory utility is not too important, the principal may provide incentives so that it is optimal for the agent to choose correct beliefs.

Our approach to boundedly rational expectations and contracting is more conservative. The agent derives her beliefs from the true data-generating process, as in the canonical

model; she just may not take into account all empirical regularities that matter for the principal's project. The misspecification in the agent's subjective model may cause her to overestimate her productivity, but, under a weak restriction, she still correctly anticipates the equilibrium distribution over output.

We also contribute to the literature on Bayesian networks/directed acyclic graphs (DAGs), which have been used extensively in the artificial intelligence literature. Pearl (2009) promotes the view that DAGs represent causal relationships and provides a broad introduction to DAGs. In economics, Spiegler (2016) and Spiegler (2017) use Bayesian networks to model agents with boundedly rational expectations. DAGs provide a general method to capture a variety of different inference errors such as reverse causation and coarseness. We build on these insights and apply them to contracting. Other recent papers use causal models to capture boundedly rational decision makers in monetary policy (Spiegler, 2020), political competition (Eliaz and Spiegler, 2020), Bayesian persuasion (Eliaz et al., 2021), and decision theory (Schenone, 2020).

The remainder of the paper is organized as follows. Section 1.2 describes our framework. In Section 1.3, we examine how a misspecification in the agent's subjective model affects the optimal contract. In Section 1.4, we state a behavioral version of the informativeness principle. In Section 1.5, we characterize when a misspecification leads to biased beliefs about the production function, and illustrate the implications of this characterization. In Section 1.6, we revisit a classic comparative static result from the canonical contracting framework. Section 1.7 concludes. Proofs and further results can be found in the appendix.

1.2 The Model

We consider a standard principal-agent problem and combine it with the Bayesian network model of boundedly rational beliefs, as introduced in Spiegler (2016).

Basic Framework. Let $A \subset \mathbb{R}$ be a finite set of actions, $Y \subset \mathbb{R}$ a finite set of outputs, and $W \subseteq \mathbb{R}^{|Y|}$ the set of possible incentive schemes. The principal proposes a contract (w, q), where $w \in W$ is the agent's wage conditional on the output $y \in Y$ and $q \in \Delta(A)$ is the probability distribution over actions that the principal wishes the agent to choose. The agent can reject or accept the contract. If she rejects it, she enjoys the outside option value $\bar{U}$, while the principal earns zero. If she accepts the contract, she chooses an action $a \in A$. The agent's personal cost of choosing a is given by a function $c(a)$. The action

stochastically influences the project's output. The agent's utility from wage w is given by the utility function $u : \mathbb{R} \to \mathbb{R}$, with $u' > 0$ and $u'' \leq 0$. When the output is y and the agent's action is a, the principal's payoff is $V = y - w(y)$ and the agent's payoff is $U = u(w(y)) - c(a)$.

Causal Structure. We model the causal structure through which the agent's action affects the output. Let $N^* = \{0, ..., n\}$ be the set of relevant variables (or nodes). This set contains the agent's action and output, but may also include other variables. A generic realization of variable i is given by $x_i \in X_i$, where X_i is a finite set that contains at least two elements. Node 0 is the agent's action ($x_0 = a$, $X_0 = A$) and node n is the output ($x_n = y$, $X_n = Y$). The state is a vector $x_{N^*} = (x_0, x_1, ..., x_n)$ and the set of all states is $X_{N^*} = \times_{i \in N^*} X_i$. For every subset $M \subseteq N^*$ and $x_{N^*} \in X_{N^*}$, we write $x_M = (x_k)_{k \in M}$.

Denote by $p(x_1, \ldots, x_n \mid a)$ the extended production function. For any action $a \in A$, it has full support over $X_1 \times ... \times X_n$. We represent its causal structure by an irreflexive, asymmetric, and acyclic binary relation R^* over N^*, and denote it by the DAG $\mathcal{R}^* = (N^*, R^*)$, see the graph on the left of Figure 1.1 for an example. For two nodes $i, j \in N^*$ one may read iR^*j as "node i impacts on node j." The set of nodes that influence i is defined, with abuse of notation, as $R^*(i) = \{j \in N^* \mid jR^*i\}$. Nothing influences the agent's action, $R^*(0) = \emptyset$. The probability distribution over states, $p(x_{N^*}) \in \Delta(X_{N^*})$, then naturally factorizes according to $\mathcal{R}^*$ via the formula

$$p(x_{N^*}) = q(x_0) \prod_{i \in N^* \setminus \{0\}} p(x_i \mid x_{R^*(i)}). \tag{1.1}$$

The "objective model" $\mathcal{R}^*$ is one of the sparsest DAGs so that $p(x_{N^*})$ factorizes according to $\mathcal{R}^*$. That is, $\mathcal{R}^*$ faithfully represents the conditional independence conditions that are satisfied by $p(x_{N^*})$; see Koski and Noble, 2009, page 39.[4]

Figure 1.1: An objective model $\mathcal{R}^*$ (left) and the agent's subjective model $\mathcal{R}$ (right).

[4]This rules out trivial cases such as when the objective distribution is consistent with the agent's subjective model (as defined below), but the agent's subjective model excludes links that are in $\mathcal{R}^*$.

Beliefs, Personal Equilibrium, and Equilibrium Contract. The agent has her own subjective model $\mathcal{R} = (N, R)$, see the graph on the right of Figure **??** for an example. We assume that $\{0, n\} \in N \subseteq N^*$ and $R(0) = \emptyset$. The assumption that the agent includes her own action and the output in her subjective model ensures that her utility is measurable with respect to her beliefs. $N \subseteq N^*$ is assumed purely for simplicity. $R(0) = \emptyset$ implies that the agent knows that she does not receive any information about other variables prior to choosing an action, and that she has correct beliefs about the marginal distribution over her own action.

Definition 1.1. *We say that $\mathcal{R}$ is misspecified if $\mathcal{R} \neq \mathcal{R}^*$, and that $\mathcal{R}$ is a simplification if $N \subset N^*$ and $R = N \times N \cap R^*$.*

A simplification is a misspecification where the agent's subjective model $\mathcal{R}$ emerges from $\mathcal{R}^*$ by dropping nodes from $\mathcal{R}^*$ and the links adjacent to them. It will receive considerable attention in this paper. However, only the results in Section 1.5 rely on the assumption that the misspecification is a simplification. Denote by $x_N = (x_i)_{i \in N}$ the state vector for the agent's subjective model and $X_N = \times_{i \in N} X_i$. The agent fits her subjective model $\mathcal{R}$ to the data generated by $p(x_{N^*})$, so her beliefs factorize according to the formula

$$p_{\mathcal{R}}(x_N) = q(x_0) \prod_{i \in N \setminus \{0\}} p(x_i \mid x_{R(i)}). \tag{1.2}$$

Thus, all the conditional independence assumptions embedded in $\mathcal{R}$ also appear in the agent's beliefs. For example, when the agent's subjective model is $\mathcal{R}$ from Figure 1.1, her beliefs factorize according to $p_{\mathcal{R}}(a, x_1, y) = q(a)p(x_1 \mid a)p(y \mid x_1)$, where $q(a)$, $p(x_1 \mid a)$ and $p(y \mid x_1)$ follow from the probability distribution $p(x_{N^*})$. Given the objective model in Figure 1.1, $p(y \mid x_1)$ will depend on q through variable 2. Hence, in contrast to the objective probabilities, the agent's beliefs about how her action influences the output may depend on q. We therefore augment notation to indicate which strategy q is used when deriving beliefs and write $p_{\mathcal{R}}(x; q)$ instead of $p_{\mathcal{R}}(x)$. For any subset $M \subset N$, the agent's belief about the marginal distribution over x_M is $p_{\mathcal{R}}(x_M; q) = \sum_{x_{N \setminus M} \in X_{N \setminus M}} p_{\mathcal{R}}(x_M, x_{N \setminus M}; q)$.

The agent follows the prescribed strategy from the contract only if it maximizes her expected utility given the wage scheme w and her subjective beliefs about the output conditional on her action, which we denote by $p_{\mathcal{R}}(y \mid a; q)$. These are computed as

$$p_{\mathcal{R}}(y \mid a; q) = \frac{p_{\mathcal{R}}(a, y; q)}{\sum_{y \in Y} p_{\mathcal{R}}(a, y; q)}. \tag{1.3}$$

To close the model, we need to specify the agent's strategy q that is used to derive these

beliefs. We adapt the personal equilibrium concept from Spiegler (2016) to our setting.

Definition 1.2. *The strategy q is a personal equilibrium at $\mathcal{R}$ and w if for all actions $a \in A$ in the support of q we have*

$$a \in \arg\max_{a'} \sum_{y \in Y} p_{\mathcal{R}}(y \mid a'; q) u(w(y)) - c(a'),$$

where $p_{\mathcal{R}}(y \mid a'; q) = \lim_{k \to \infty} p_{\mathcal{R}}(y \mid a'; q^k)$ for all actions $a' \in A$ and a sequence $q^k \to q$ of fully mixed strategy profiles.

With the full support assumption, a fully mixed action profile ensures that all conditional probabilities are well-defined. The definition requires that equilibrium beliefs are the limit of a sequence of fully mixed profiles. The equilibrium beliefs are independent of the sequence of fully mixed strategies used to approximate them, and a personal equilibrium always exists in our framework; see Appendix 1.8.1. We call a contract (w, q) an "equilibrium contract" if q is a personal equilibrium at $\mathcal{R}$ and w. An optimal equilibrium contract is an equilibrium contract that maximizes the principal's expected payoff. For convenience, we denote beliefs by $p_{\mathcal{R}}(y \mid a; a^*)$ when a pure action a^* is implemented, and $p_{\mathcal{R}}(y \mid a; \alpha)$ with $q(a = 1) = \alpha$ when we have a binary action set $A = \{0, 1\}$.

Instead of considering a personal equilibrium, we could in principle assume that the agent derives beliefs from some arbitrary joint probability distribution $\hat{p}(x_N)$. In this case, we would have a model with exogenously fixed biased beliefs $\hat{p}(y \mid a)$. The personal equilibrium definition imposes restrictions on the agent's beliefs: Through the factorization in equation (1.2), they must respect the agent's strategy q and the extended production function. One interpretation is that the agent is experienced and thus has data on how her action impacts on the variables in her subjective model. An alternative interpretation is that there are (or have been) many other agents in the organization who exchange data with their new colleague to which she can fit her subjective model. One might suppose that the agent estimates the distribution over the output separately for each available action. This is however not what happens in this model. Instead, the agent "pools" the data from different actions when she estimates the conditional probabilities for variables that (according to her subjective model) are not directly influenced by her action. We return to this discussion at the end of Subsection 3.2.

1.3 The Optimal Equilibrium Contract

In this section, we study the properties of the optimal equilibrium contract for a given extended production function $p(x_1, ..., x_n \mid a)$ and subjective model $\mathcal{R}$. If (w^*, q^*) is an optimal equilibrium contract, then w^*, q^* solve the maximization problem

$$\max_{w \in W, q \in \Delta(A)} \sum_{a \in A} \sum_{y \in Y} q(a) p(y \mid a)(y - w(y)) \tag{1.4}$$

subject to the constraints

$$q \in \Delta(A) \quad \textit{is a personal equilibrium at } \mathcal{R} \textit{ and } w, \tag{IC}$$

$$\sum_{a' \in A} \sum_{y \in Y} q(a')[p_{\mathcal{R}}(y \mid a'; q) u(w(y)) - c(a')] \geq \bar{U}. \tag{PC}$$

When the agent's subjective model $\mathcal{R}$ equals the objective model $\mathcal{R}^*$, the problem collapses to the canonical principal-agent problem, and can be solved as suggested by Grossman and Hart (1983). We first find for each pure action $a \in A$ the wage scheme w that implements this action at lowest possible cost. Then we choose the action-incentive scheme combination that maximizes the principal's profit. If the agent's subjective model $\mathcal{R}$ differs from the objective model $\mathcal{R}^*$, we find the optimal equilibrium contract by applying the same procedure. However, since the agent's beliefs $p_{\mathcal{R}}(y \mid a; q)$ may depend on the implemented strategy q, the first step has to be done for all pure and mixed strategies $q \in \Delta(A)$.

Suppose the agent is risk-averse with unlimited liability, and the principal implements a (possibly mixed) strategy q. The Kuhn-Tucker conditions for the principal's problem are then necessary and sufficient for an optimum. Choose any action a in the support of q. The optimal incentive scheme is then characterized by the first-order condition

$$\frac{1}{u'(w(y))} = \frac{p_{\mathcal{R}}(y; q)}{p(y)} \left[\mu + \sum_{a' \in A} \lambda_{a'} \frac{p_{\mathcal{R}}(y \mid a; q) - p_{\mathcal{R}}(y \mid a'; q)}{p_{\mathcal{R}}(y; q)} \right] \tag{1.5}$$

for all $y \in Y$, where μ and $\lambda_{a'}$ are the usual Lagrange multipliers for the participation and incentive compatibility constraint, respectively. Equation (1.5) allows us to disentangle how a misspecification in $\mathcal{R}$ may change the contracting problem. First, the PC is affected when the agent holds biased beliefs about the equilibrium distribution over output; see the first term on the right of equation (1.5). In Subsection 1.3.1, we state a sufficient condition on $\mathcal{R}$ so that this belief is unbiased. Second, the IC may be affected. Suppose the principal implements a pure action a and $p_{\mathcal{R}}(y; a) = p(y)$. The ratio in the squared brackets then

becomes $1 - \frac{p_{\mathcal{R}}(y|a';a)}{p_{\mathcal{R}}(y|a;a)}$, in which case the optimal incentive scheme depends on a likelihood ratio as in the canonical framework. Any difference between the contracts under the objective and subjective model is then driven by differences between the corresponding likelihood ratios. In Subsection 1.3.2, we examine in an example how these differences may affect the optimal equilibrium contract.

1.3.1 Correct Expectations on the Equilibrium Path

We use a Bayesian network result from Spiegler (2017) that characterizes under what circumstances the agent's beliefs about the equilibrium output distribution are correct, so that $p_{\mathcal{R}}(y; q) = p(y)$ for all $q \in \Delta(A)$. To this end, we introduce a few definitions. A v-collider is a triple of nodes (i, j, k) such that iRj, kRj and there is no link between i and k (neither iRk nor kRi is in R). The set of v-colliders of a DAG is called its v-structure. A DAG is called perfect if it has an empty v-structure. A subset of nodes $M \subset N$ is a clique in $\mathcal{R} = (N, R)$ if iRj or jRi for any two nodes $i, j \in M$. For example, in the DAG $\mathcal{R}^*$ from Figure **??**, the set $M = \{1, 3, 4\}$ is a clique, while the set $M' = \{2, 3, 4\}$ is not. Each node is a clique in itself, so the output node n is a clique. The following result essentially restates Proposition 2 from Spiegler (2017).

Proposition 1.1 (Equilibrium Beliefs)**.** *If the agent's model $\mathcal{R} = (R, N)$ is perfect, her equilibrium beliefs satisfy $p_{\mathcal{R}}(x_M; q) = p(x_M)$ for all $q \in \Delta(A)$ and any clique $M \subset N$.*

If the agent's subjective model $\mathcal{R}$ is perfect, then, in a personal equilibrium, the agent correctly anticipates the marginal distribution over each variable in her model, and also the joint distribution over variables in cliques. The intuition behind this result is that perfectness excludes biased estimates due to neglect of correlation. Imagine two variables i, j that influence a third variable k. Suppose that i and j are correlated, and that the agent treats them as uncorrelated. Through the application of the factorization formula (1.2), the agent may then obtain a biased estimate of the marginal distribution over k. Perfectness implies that the agent always checks for correlations between two variables i, j when, according to her subjective model, they influence a third variable k. We obtain two useful corollaries from Proposition 1.1.

Corollary 1.1. *If the agent's model $\mathcal{R} = (R, N)$ is perfect and her equilibrium strategy is a pure action a^*, her equilibrium beliefs satisfy $p_{\mathcal{R}}(x_M \mid a^*; a^*) = p(x_M \mid a^*)$ for every clique $M \subset N$.*

If the equilibrium contract implements a pure strategy a^*, the agent's belief about the joint distribution of any clique M conditional on her equilibrium strategy is correct. Corollary 1.1 is in general not true if the equilibrium contract implements a mixed strategy q^*. While the agent still gets the marginal equilibrium distribution over each variable right, her beliefs may also exhibit $p_{\mathcal{R}}(x_i \mid a'; q^*) \neq p(x_i \mid a')$ for an action a' in the support of q^*. Thus, the agent's expected utility conditional on a' may be biased, $\mathbb{E}_{\mathcal{R}}[u(w(y)) \mid a'; q^*] \neq \mathbb{E}[u(w(y)) \mid a']$.

The second direct implication of Proposition 1.1 is the following result.

Corollary 1.2. *Suppose (w, q) is an equilibrium contract. If $\mathcal{R} = (R, N)$ is perfect, the PC is satisfied at this contract if and only if this is also the case under the objective model $\mathcal{R}^*$.*

To see why Corollary 1.2 is true recall that every single node is a clique. Hence, Proposition 1.1 implies $p(y) = p_{\mathcal{R}}(y; q) = \sum_{a \in A} q(a) p_{\mathcal{R}}(y \mid a; q)$. If $\mathcal{R}$ is perfect, the incentive scheme therefore has to satisfy the same participation constraint as under the objective model. Thus, an agent with a misspecified – but perfect – model cannot be exploited. Throughout the paper, we will assume that $\mathcal{R}$ is perfect. As we see next, a perfect $\mathcal{R}$ does not imply that the principal cannot benefit from the agent's misperception.

1.3.2 Incentive Effects

We examine how a misspecification in the agent's subjective model $\mathcal{R}$ can change the equilibrium contract. We do this in the context of the marketer example from the introduction. Figure 1.2 shows the objective model $\mathcal{R}^*$ and the agent's subjective model $\mathcal{R}$.

Figure 1.2: Objective model $\mathcal{R}^*$ (left) and subjective model $\mathcal{R}$ (right) in the marketer example.

Since the marketer believes that her action only affects output through the information channel (node 1), her subjective model $\mathcal{R}$ is perfect. By Corollary 1.2, only the incentive compatibility constraint can then be affected by the misspecification. We analyze a simple

setting with two effort levels $a \in \{0,1\}$, two output levels $y \in \{y_L, y_H\}$ with $y_H > y_L$, and cost $c(1) = c > c(0) = 0$. The probability of output y_H increases in the agent's effort. Node 1 is the level of consumer information. It can be low ($x_1 = 0$) or high ($x_1 = 1$). Node 2 is the firm's reputation, which can be bad ($x_2 = 0$) or good ($x_2 = 1$). The subjective model $\mathcal{R}$ captures that the agent does not take reputation into account. For the objective probability distribution, we use the parametrization $p(x_i = 1 \mid x_0) = \beta_i + \beta_{0i}x_0$ for $i \in \{1,2\}$ and $p(y_H \mid x_1, x_2) = \beta_3 + \beta_{13}x_1 + \beta_{23}x_2$. Making cold-calls increases consumer information, $\beta_{01} > 0$, and decreases reputation, $\beta_{02} < 0$; consumer information x_1 and reputation x_2 both have a positive influence on sales, $\beta_{13} > 0$ and $\beta_{23} > 0$. We obtain the following result.

Proposition 1.2 (Marketer Example)**.** *Consider the marketer example of this subsection.*

(a) The simplification in the agent's subjective model $\mathcal{R}$ relaxes the IC for $\alpha = 1$.

(b) The optimal equilibrium contract implements $\alpha \in \{0,1\}$. If and only if effort costs c are small enough, the optimal equilibrium contract implements $\alpha = 1$ and the principal strictly benefits from the simplification in the agent's subjective model $\mathcal{R}$.

Before we prove this result, we explain the intuition behind it and its implications. First, consider statement (a). When the principal implements $\alpha = 1$, the agent overestimates the drop in expected output when she exerts low instead of high effort. According to her subjective model $\mathcal{R}$, the only effect of her action on the output occurs through consumer information x_1. She does not take into account that a deviation to low effort would also have a positive effect on expected reputation, which translates into a positive effect on expected output. Formally, the IC under the objective model $\mathcal{R}^*$ is

$$[\beta_{01}\beta_{13} + \beta_{02}\beta_{23}]\ (u(w(y_H)) - u(w(y_L))) - c \geq 0. \tag{1.6}$$

The term in squared brackets is the effect of effort on output and contains the consumer information channel $\beta_{01}\beta_{13}$ and the reputation channel $\beta_{02}\beta_{23}$. Under the subjective model $\mathcal{R}$, this second channel is missing. When the agent calibrates her model, she correctly estimates the impact of her action on the distribution of consumer information. However, when she estimates the impact of consumer information on sales, it is as if she suffers from omitted variable bias, and her estimate will depend on the implemented strategy α. Hence, the perceived effect of action on sales – and therefore also the IC – depends on α. In the proof of Proposition 1.2 , we derive the IC for all $\alpha \in [0,1]$. For $\alpha = 1$ the IC

becomes

$$\beta_{01}\beta_{13}\ (u(w(y_H)) - u(w(y_L))) \geq c. \tag{1.7}$$

Since the effect of effort on reputation β_{02} is negative, the simplification in $\mathcal{R}$ relaxes the IC. As long as $\alpha \in (0,1)$, the reputation effect is partly reflected in $p(y_H \mid x_1)$. The extent of this depends on α since α affects the correlation between consumer information and reputation. A higher correlation between consumer information and reputation would mitigate some of the effect of the agent's misperception.

Next, consider statement (b). The observation that the principal implements a pure strategy would be trivial in the canonical framework with rational expectations. This is not the case here as the agent's perceived effect of effort on output $p_{\mathcal{R}}(y_H \mid a = 1;\alpha) - p_{\mathcal{R}}(y_H \mid a = 0;\alpha)$ may vary non-monotonically in α. In the present setting, the perceived effect of effort on output is maximal at $\alpha = 1$, so that there is no reason for the principal to implement a mixed strategy. At the end of this subsection, we present an example where the unique optimal equilibrium contract indeed implements a mixed strategy $\alpha \in (0,1)$.

Importantly, if the agent chooses a pure strategy, then, by Corollary 1.1 and the fact that $\mathcal{R}$ is perfect, she correctly anticipates the joint distribution over all variables in $\mathcal{R}$ conditional on her equilibrium action. Thus, in the data that the agent gets under the optimal equilibrium contract, there are no informational cues which could alarm her about a misspecification in her subjective model. This is a crucial difference between the present framework and models where beliefs about outcomes are biased for equilibrium actions.

Finally, the last part of statement (b) spells out that the principal strictly benefits from the simplification in $\mathcal{R}$ when effort costs are small enough so that it is profitable to implement high effort. For a range of effort costs c, the principal implements low effort when the agent has rational expectations, but high effort if her subjective model is $\mathcal{R}$. This is of course not true in general. For example, if the agent's action has a positive effect on reputation, $\beta_{02} > 0$, the simplification in $\mathcal{R}$ tightens the IC for $\alpha = 1$ as the agent does not take all positive effects of her action on output into account.

To illustrate our approach, we present the proof of Proposition 1.2.

Proof of Proposition 1.2. We first derive $p_{\mathcal{R}}(y_H \mid a;\alpha)$ for a given mixed equilibrium strategy $\alpha \in (0,1)$. The agent's equilibrium belief about the joint probability distribution of the variables in $\mathcal{R}$ is given by $p_{\mathcal{R}}(a, x_1, y) = q(a)p(x_1 \mid a)p(y \mid x_1)$. Since node 0 and

node 1 form a clique and $\mathcal{R}$ is perfect, the agent's belief about the joint probability distribution of a and x_1 is correct. Hence, $p(x_1 \mid a)$ is independent of α and we have $p(x_1 = 1 \mid a) = \beta_1 + \beta_{01}a$. However, $p(y \mid x_1)$ depends on α since the distribution over y also depends on x_2. To get $p(y \mid x_1)$, we first derive $p(x_2 = 1 \mid x_1)$, i.e., the probability that $x_2 = 1$ given that value x_1 is observed at node 1 when the agent's equilibrium action is α. We calculate

$$p(x_2 = 1 \mid x_1 = 1) = \frac{\alpha(\beta_1 + \beta_{01})(\beta_2 + \beta_{02}) + (1-\alpha)\beta_1\beta_2}{\beta_1 + \alpha\beta_{01}}, \tag{1.8}$$

$$p(x_2 = 1 \mid x_1 = 0) = \frac{\alpha(1 - \beta_1 - \beta_{01})(\beta_2 + \beta_{02}) + (1-\alpha)(1-\beta_1)\beta_2}{1 - \beta_1 - \alpha\beta_{01}}. \tag{1.9}$$

With this we can calculate the equilibrium probability that output y_H realizes after observing $x_1 = 1$ and $x_1 = 0$, respectively:

$$p(y_H \mid x_1 = 1) = \beta_3 + \beta_{13} + \frac{\alpha(\beta_1 + \beta_{01})(\beta_2 + \beta_{02}) + (1-\alpha)\beta_1\beta_2}{\beta_1 + \alpha\beta_{01}}\beta_{23}, \tag{1.10}$$

$$p(y_H \mid x_1 = 0) = \beta_3 + \frac{\alpha(1 - \beta_1 - \beta_{01})(\beta_2 + \beta_{02}) + (1-\alpha)(1-\beta_1)\beta_2}{1 - \beta_1 - \alpha\beta_{01}}\beta_{23}. \tag{1.11}$$

From $p_{\mathcal{R}}(a, x_1, y)$ we can now calculate the agent's subjective probability of a high output after high and low effort, respectively:

$$p_{\mathcal{R}}(y_H \mid a = 1; \alpha) = (\beta_1 + \beta_{01})p(y_H \mid x_1 = 1) + (1 - \beta_1 - \beta_{01})p(y_H \mid x_1 = 0), \tag{1.12}$$

$$p_{\mathcal{R}}(y_H \mid a = 0; \alpha) = \beta_1 p(y_H \mid x_1 = 1) + (1 - \beta_1)p(y_H \mid x_1 = 0). \tag{1.13}$$

We then use these terms to compute the IC for $\alpha \in (0, 1)$,

$$\left[p_{\mathcal{R}}(y_H \mid a = 1; \alpha) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha)\right] \left(u(w(y_H)) - u(w(y_L))\right) = 0. \tag{1.14}$$

By taking the limit for $\alpha \to 1$, we obtain the IC for $\alpha = 1$, which is the inequality in (1.7). Since $\beta_{02} < 0$, this completes the proof of statement (a). To prove statement (b), note first that both IC and PC must be binding at the optimal equilibrium contract. Simple calculations show that $\beta_{01}, \beta_{13}, \beta_{23} > 0$ and $\beta_{02} < 0$ imply

$$p_{\mathcal{R}}(y_H \mid a = 1; \alpha) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha) \leq \beta_{01}\beta_{13} \tag{1.15}$$

for all $\alpha \in (0, 1]$; that is, when the agent exerts high effort with positive probability, her perceived effect of effort on output is largest at $\alpha = 1$. The principal then cannot gain from implementing a mixed strategy. Finally, given that the optimal equilibrium contract implements either $\alpha = 0$ or $\alpha = 1$, the last part of statement (b) follows from a simple

comparison of expected profits under the equilibrium contracts that implement these two actions. □

Mixed strategy example. We show by example that it is not always optimal for the principal to implement a pure strategy. Consider again the marketer example. Assume that the agent is risk-neutral, protected by limited liability so that $w \geq 0$, her outside option value is zero, and $y_L = 0$. Suppose payoff parameters are such that the principal optimally implements some $\alpha > 0$. Standard arguments show that $w(y_L) = 0$, and that $w(y_H)$ is chosen so that the IC in (1.14) is satisfied. The principal's expected payoff from this contract is then

$$\mathbb{E}[V] = [\alpha p(y_H \mid a = 1) + (1 - \alpha)p(y_H \mid a = 0)] \left(y_H - \frac{c}{\Delta_{\mathcal{R}}(\alpha)} \right), \tag{1.16}$$

where $\Delta_{\mathcal{R}}(\alpha) = p_{\mathcal{R}}(y_H \mid a = 1; \alpha) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha)$ is the agent's perceived effect of effort on output. The slope of $\Delta_{\mathcal{R}}(\alpha)$ at $\alpha = 1$ is

$$\left. \frac{d\Delta_{\mathcal{R}}(\alpha)}{d\alpha} \right|_{\alpha=1} = \beta_{01}\beta_{02}\beta_{23} \left(\frac{\beta_1}{\beta_1 + \beta_{01}} - \frac{1 - \beta_1}{1 - \beta_1 - \beta_{01}} \right). \tag{1.17}$$

Let the agent's action have a positive impact on both consumer information and reputation, $\beta_{01} > 0$ and $\beta_{02} > 0$. Then for $\beta_{01} \to 1 - \beta_1$ the slope in (1.17) converges to minus infinity. Hence, if all else equal β_{01} is sufficiently close to $1 - \beta_1$, then, starting from $\alpha = 1$, a small reduction in α reduces $w(y_H)$, and in terms of profits, this reduction overcompensates the smaller probability of high output. The optimal equilibrium contract then implements a mixed strategy. Thus, when the agent is induced to switch between periods of working hard and periods of shirking, her effort appears to her as particularly important for the final output.

Of course, when the agent chooses a mixed strategy, then the data generated in equilibrium would suffice to identify the real effect of effort on output. For this, the agent would have to analyze the data like an experimentalist and compare the average output under high and low effort, respectively. However, according to her subjective model, this "test" is unnecessary, and she therefore saves herself the trouble of performing it. Thus, one interpretation for the mixed strategy equilibrium is that the agent does not use her data effectively to correctly derive the effect of her effort on output.

1.3.3 Justifiability

In our framework, the agent has a fully specified model that makes predictions about outcomes for all actions $a \in A$. A natural question is then whether the optimal equilibrium contract is also optimal for the principal when evaluated from the agent's (potentially biased) perspective. If according to her subjective beliefs the principal should have offered another contract, the agent may suspect that her subjective model $\mathcal{R}$ is not correct.[5] We call this refinement "justifiability." It has first been defined in the unawareness literature by Filiz-Ozbay (2012). We can conveniently adapt it to our framework. In the following definition, we distinguish between "justifiability" and "partial justifiability."

Definition 1.3. *An equilibrium contract (w^*, q^*) is justifiable at $\mathcal{R}$ if w^*, q^* solve the maximization problem*

$$\max_{w \in W, q \in \Delta(A)} \sum_{a \in A} \sum_{y \in Y} q(a) p_{\mathcal{R}}(y \mid a; q^*)(y - w(y))$$

subject to the constraints that, for all a in the support of q, we have

$$a \in \arg_{a' \in A} \max \sum_{y \in Y} p_{\mathcal{R}}(y \mid a'; q^*) u(w(y)) - c(a'), \text{ and}$$
$$\sum_{a \in A} \sum_{y \in Y} q(a)[p_{\mathcal{R}}(y \mid a; q^*) u(w(y)) - c(a)] \geq \bar{U}.$$

An equilibrium contract (w^*, q^*) is partially justifiable at $\mathcal{R}$ if w^* is a solution to this maximization problem when $q = q^*$ is given.

An equilibrium contract (w^*, q^*) is justifiable if the choice of the incentive scheme w^* and the implemented strategy q^* maximizes the principal's expected payoff when evaluated according to the agent's beliefs $p_{\mathcal{R}}(y \mid a; q^*)$. It is partially justifiable if the incentive scheme w^* maximizes the principal's expected payoff, when evaluated according to the agent's beliefs, given that the principal wants to implement strategy q^*. Partial justifiability is a weaker refinement where the agent does not doubt her subjective model if at least the incentive scheme appears to be optimal for the principal. We examine under what circumstances an optimal equilibrium contract is (partially) justifiable, and obtain this result:

[5] We do not model how in this case the agent adjusts her subjective model. One alternative is that, after becoming suspicious, she looks at the production process more closely and discovers the objective model $\mathcal{R}^*$.

Proposition 1.3 (Justifiability)**.** *Let (w^*, q^*) be an optimal equilibrium contract. If we have $p_{\mathcal{R}}(y; q) = p(y)$ for all $q \in \Delta(A)$, the following statements hold:*

(a) This contract is partially justifiable at $\mathcal{R}$.

(b) If A, Y are binary sets, q^* is a pure strategy, and the principal strictly prefers this contract to the optimal contract under the objective model $\mathcal{R}^*$, it is justifiable at $\mathcal{R}$.

The proof of this result is in Appendix 1.8.2. The first part of Proposition 1.3 states that an optimal equilibrium contract is partially justifiable if the agent has correct expectations on the equilibrium path. In this case, the maximization problem in (1.4) and that in Definition 1.3 are identical for a given strategy q^*. The optimal incentive scheme that implements q^* then also appears to the agent as optimal for the principal. Thus, by Proposition 1.1, if the agent's subjective model $\mathcal{R}$ is perfect, the optimal equilibrium contract is partially justifiable at $\mathcal{R}$.

This is a significant difference to a framework where the agent's beliefs $\hat{p}(y \mid a)$ are exogenously fixed. The optimal contract in such a framework may not be partially justifiable since it may contain a bet that, from the agent's perspective, is not optimal for the principal. To illustrate, consider the two-actions-two-outcomes example from the previous subsection. Suppose that the principal implements high effort $\alpha = 1$, and that the agent's beliefs are biased so that $\hat{p}(y_H \mid a = 1) > p(y_H \mid a = 1)$ and $\hat{p}(y_H \mid a = 0) = p(y_H \mid a = 0)$. Now let effort costs c converge to zero. Under rational expectations, this would imply that the optimal contract converges to a fixed-wage contract. In contrast, under biased beliefs, the optimal contract remains bounded away from fixed wages: To exploit the agent's bias, it pays more to her after output y_H and less after output y_L. However, from the agent's perspective, an incentive scheme that is close to fixed wages would be optimal. Thus, from her perspective, the offered incentive scheme cannot be optimal for the principal.

To prove justifiability, we additionally have to show that, according to the agent's beliefs, the principal cannot benefit from implementing a different action. Unfortunately, it is then no longer possible to derive a general statement. If an equilibrium contract is optimal for the principal, this does not imply that it is justifiable, even if the agent has correct expectations on the equilibrium path. Justifiability then has to be proven for each case individually.

The second part of Proposition 1.3 states sufficient (but not necessary) conditions for justifiability for a relevant special case. In a two-actions-two-outcomes setting, an optimal

equilibrium contract is justifiable if it implements a pure strategy and the principal strictly benefits from the agent's misperception (as in the marketer example of the previous subsection). The requirement of a pure strategy is crucial here. Consider the mixed strategy example from the previous subsection where the principal implements $\alpha \in (0,1)$ to alter the agent's sense for the importance of her effort. From the agent's perspective, this does not make sense. According to her, it would be optimal for the principal to implement high effort with certainty. Note that she is indifferent between high and low effort, so (in her mind) the incentive scheme can remain the same. Thus, the optimal equilibrium contract in the mixed strategy example is not justifiable.

1.4 The Informativeness Principle

An important question in contract theory is on which information the principal should condition the agent's wage. For a setting with a risk-averse agent who has unlimited liability, the informativeness principle states that the optimal contract conditions on an additional variable z if and only if it is informative about the agent's effort, i.e., if and only if the likelihood ratio $\frac{p(y,z|a')}{p(y,z|a)}$ varies in z for some y.[6] In this section, we derive a version of the informativeness principle that allows for boundedly rational agents. To this end, we exploit the fact that an agent with biased subjective beliefs may still have correct expectations about the joint distribution of the contractible variables in equilibrium. We then apply our version of the informativeness principle to provide a rationale for why in executive compensation contracts peer-performance is mostly not used so that CEOs are rewarded for windfall gains.

The original version of the informativeness principle may no longer hold when the agent's subjective model $\mathcal{R}$ is misspecified. Consider the marketer example from Subsection 1.3.2 and assume that the principal can also condition the agent's wage on consumer information x_1. If the agent had rational expectations, the optimal wage scheme would condition both on consumer information x_1 and sales x_3 since neither variable is a sufficient statistic of the other (to avoid confusion below, we here use x_3 instead of y). However, according to the agent's subjective model $\mathcal{R}$, sales x_3 are just a noisy signal of consumer information x_1. Therefore, the optimal equilibrium contract only conditions on x_1 and appears as "incomplete."[7]

[6] Whether this result holds or not depends on the formal details of the contracting problem; see Chaigneau et al. (2019) for a recent discussion and a further extension of the informativeness principle.

[7] A further interesting trade-off can be observed here. Recall from the marketer example that when

We can generalize this finding and obtain a version of the informativeness principle that allows for misspecified subjective models $\mathcal{R}$. To get this statement, we assume that the agent's subjective model is such that she correctly anticipates the joint distribution over the two contractible variables y and z. Recall from Proposition 1.1 that this is the case if $\mathcal{R}$ is perfect and there is a link between y and z in $\mathcal{R}$ (so that they form a clique).

Proposition 1.4 (Informativeness Principle)**.** *Suppose the agent is risk-averse and has unlimited liability. Let y and z be two contractible variables that are both part of the agent's subjective model $\mathcal{R}$. If $p_{\mathcal{R}}(z, y; q) = p(z, y)$ for all $q \in \Delta(A)$, the following statements hold:*

(a) Suppose that $a \in \{0, 1\}$ and $c(1) > c(0)$. The equilibrium contract that implements $\alpha = 1$ at lowest cost to the principal does not condition on z if and only if for all triples a, y, z we have $p_{\mathcal{R}}(z \mid y, a; \alpha = 1) = p_{\mathcal{R}}(z \mid y; \alpha = 1)$.

(b) If for all $q \in \Delta(A)$ and all triples a, y, z we have $p_{\mathcal{R}}(z \mid y, a; q) = p_{\mathcal{R}}(z \mid y; q)$, the optimal equilibrium contract does not condition on z.

The proof of Proposition 1.4 is in Appendix 1.8.3. We provide an interpretation of this result and explain its implications. First, the condition $p_{\mathcal{R}}(z \mid y, a; q) = p_{\mathcal{R}}(z \mid y; q)$ for all $q \in \Delta(A)$ and all triples a, y, z indicates that, in the agent's mind, variable z is independent of her action conditional on variable y (regardless of the implemented action). If this condition is satisfied, the agent believes that z does not contain any information about her action that is not already in y. However, this condition alone does not imply that the optimal equilibrium contract does not condition the agent's wage on z. In addition, the agent's subjective belief about the joint equilibrium distribution of y and z needs to be correct. Otherwise, the principal may want to exploit the agent's biased perception of this distribution, and condition on z even if the agent thinks that z is uninformative about her action given y. This is equivalent to betting when two individuals have different prior beliefs about future events (as pointed out in the previous subsection, such a contract would also not be justifiable).

An interesting special case emerges when the agent believes that z is independent of all other variables. If y and z are independent in the objective model, the optimal equilibrium contract would not condition on z (even if y and z are not independent conditional on

the contract only conditions on sales x_3, the agent with subjective model $\mathcal{R}$ is control optimistic, which relaxes the IC. In contrast, when the contract only conditions on consumer information x_1, the agent has correct expectations about her expected payoff under alternative actions, so the IC is unaffected by the misspecification in $\mathcal{R}$. Nevertheless, it is optimal for the principal to condition the agent's wage only on x_1 as it is a more precise signal about her effort than sales x_3, that is, the informativeness effect dominates the incentive effect from the misspecification since $p(x_1 = 1 \mid a = 1) - p(x_1 = 1 \mid a = 0) = \beta_{01} > \beta_{01}\beta_{13} = p_{\mathcal{R}}(y_H \mid a = 1; \alpha = 1) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha = 1)$.

a). From the agent's perspective that would only introduce noise to the wage scheme. However, if z and y are correlated, the requirements of Proposition 1.4 are no longer satisfied, and the optimal equilibrium contract may imply a bet on the joint realization of y and z.

Second, Proposition 1.4 consists of two statements. Statement (a) is the informativeness principle for the case of binary action spaces. It is very similar to the original version: The statement implies that the optimal equilibrium contract that implements $\alpha = 1$ conditions on z if and only if the likelihood ratio $\frac{p_{\mathcal{R}}(y,z|a=0;\alpha=1)}{p_{\mathcal{R}}(y,z|a=1;\alpha=1)}$ varies in z for some y. Statement (b) for general finite action spaces is weaker since the additional information embedded in z may, according to the agent's subjective beliefs, only affect non-binding ICs.[8]

Third, observe that Proposition 1.4 does not impose any further assumptions on the agent's subjective model $\mathcal{R}$. It therefore applies to all settings in which the agent's beliefs satisfy the conditions outlined in the proposition. Importantly, we can state sufficient conditions on $\mathcal{R}$ so that the agent's beliefs satisfy the conditional independence assumption. The Bayesian network literature establishes "d-separation" as a convenient tool to check conditional independence of two sets of variables in a model $\mathcal{R}$; we describe it in a supplementary appendix.

Fourth, our Bayesian network framework allows for a causal interpretation of the informativeness principle. The optimal equilibrium contract conditions on both y and z if the agent's action has partially independent effects on these two variables according to $\mathcal{R}$. It does not condition on z if, according to $\mathcal{R}$, variable z is a consequence of y. In this case, the optimal contract conditions on the variable that is "causally closer" to the agent's action.

As an application, we consider a setting in which the principal can condition the agent's wage both on her output $y \in \{y_L, y_H\}$ and on her relative performance $z \in \{-1, 0, 1\}$. The latter variable captures, for example, how the stock price of the company compares to that of the company's rivals. There is a common shock $x_1 \in \{0, 1\}$, e.g., the state of the economy, that positively affects both own output y and the rivals' output $x_3 \in \{y_L, y_H\}$. Through competition, output y has a negative effect on the rivals' output x_3 (e.g., if y is high, the rivals' output tends to be smaller since consumers prefer the product of the agent's firm). The objective model $\mathcal{R}^*$ on the left in Figure 1.3 illustrates this setting.

[8] This is a general issue of the informativeness principle and not specific to our framework.

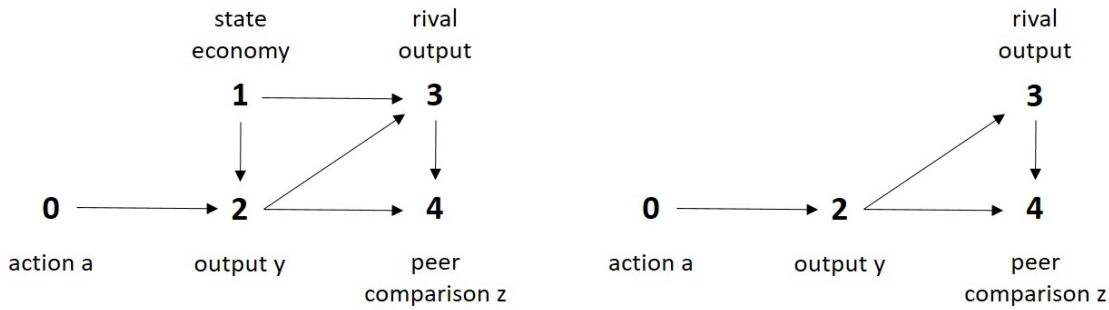

Figure 1.3: Objective model $\mathcal{R}^*$ (left) and subjective model $\mathcal{R}$ (right) in the peer-comparison example.

Under the objective model $\mathcal{R}^*$, the optimal equilibrium contract that implements high effort would, at any generic parametrization, condition the agent's wage both on output and relative performance. This can be established by visually inspecting $\mathcal{R}^*$ using d-separation.[9] The intuition is as follows: Suppose we know the agent's output y. Then information about the agent's action a provides additional information about the state of the economy x_1, and hence also additional information about peer performance z. Hence, a and z are not independent conditional on y in $\mathcal{R}^*$.

Now suppose that the agent does not take the common shock x_1 into account so that her subjective model is given by $\mathcal{R}$ on the right of Figure 1.3. Since $\mathcal{R}$ is perfect and the variables y and z are linked in $\mathcal{R}$, the agent correctly anticipates the equilibrium distribution over the two variables. Moreover, if we know the output y, then, according to $\mathcal{R}$, the agent's action contains no further information about z (one can formally show this using d-separation). Proposition 1.4 then implies that the optimal equilibrium contract that implements $\alpha = 1$ only conditions on the agent's own output y. It is therefore incomplete and rewards the agent for windfall gains that come from good states of the economy. In the agent's mind, her relative performance is only a noisy signal of her own output. Hence conditioning her wage on relative performance would only increase the agent's exposure to risk and hence implementation costs.

Many actual compensation contracts indeed do not make use of peer-performance and reward executives for windfall gains. Bertrand and Mullainathan (2001) and Bebchuk and Fried (2004) discuss this phenomenon and possible explanations. A popular explanation is that executives use their influence over the board of directors to alter their compensation, which then happens to increase in windfall gains. However, this theory cannot explain the

[9]The "usual" way to see this is to consider a particular parametrization. Consider our linear specification with binary outcomes at all variables except z. For z we assume that $p(z = 1 \mid y > x_3) \approx 1$, $p(z = 0 \mid y = x_3) \approx 1$, and $p(z = -1 \mid y < x_3) \approx 1$. If the influence of y on x_3 is small enough, the optimal contract that implements high effort conditions on both variables, and the agent's wage increases in both y and z.

inefficient risk allocation. In contrast, model misspecification can account for inefficient risk allocation. For example, the manager's model is misspecified as in the application if she attributes the output to her action alone, or if she ignores the statistical implications of common shocks and therefore evaluates peer-performance as uninformative about her own action.

1.5 Behavioral Rationality

We learned in Section 1.3 that a simplification in the agent's subjective model may affect the incentive compatibility constraint. However, does a simplification in $\mathcal{R}$ automatically imply that the agent's beliefs are biased? In this section, we show that the answer is negative. The agent may correctly anticipate the true production function even when her subjective model $\mathcal{R}$ omits variables from $\mathcal{R}^*$. When this statement holds for any parametrization of the extended production function that factorizes[10] according to $\mathcal{R}^*$, we say that the agent is "behaviorally rational." We state the formal definition.

Definition 1.4. *An agent with subjective model $\mathcal{R}$ is behaviorally rational if, at any probability distribution $p \in \Delta(X_{N^*})$ that factorizes according to $\mathcal{R}^*$, we have $p_{\mathcal{R}}(y \mid a; q) = p(y \mid a)$ for all $a \in A$ and $q \in \Delta(A)$.*

For a given objective model $\mathcal{R}^*$ we can characterize when the agent is behaviorally rational, provided that the misspecification is a simplification. This restriction is useful as it implies that $\mathcal{R}^*$ and the set of nodes in the agent's subjective model N fully characterize $\mathcal{R}$. We will see that two extended production functions – which involve the same set of nodes N^* and may give rise to the same production function $p(y \mid a)$ – can differ in the extent to which simplifications affect the agent's beliefs about $p(y \mid a)$. This extent depends on the "channels" in $\mathcal{R}^*$ through which the agent's action affects the output. Intuitively, they describe the agent's role in the organization, that is, which components or behaviors of others the agent affects directly or indirectly through her action. In Subsection 1.5.1, we motivate this interpretation in an example where the agent's job determines the scope for biased beliefs and control optimism. In Subsection 1.5.2, we characterize when the agent is behaviorally rational and generalize the main findings from Subsection 1.5.1.

[10] In this section, we deviate from our earlier assumption that $p(x_{N^*})$ does not contain any additional conditional independence assumptions compared to $\mathcal{R}^*$. This allows us to use results and techniques from the Bayesian network literature. Importantly, if the agent is behaviorally rational in the current setting, she is also behaviorally rational under the earlier assumption.

1.5.1 The Agent's Job and the Scope for Control Optimism

We examine the interaction between the agent's job, model misspecification, and incentives. Let the agent first work as an ordinary marketer whose job is to increase sales. This time, making cold-calls is not part of her job. Her effort only has a (positive) effect on consumer information, for example, through informative advertising. Nevertheless, there is a group of employees engaged in telemarketing. Their effort – making cold-calls – impacts on consumer information and the firm's reputation in the usual manner. The objective model $\mathcal{R}^*$ on the left of Figure 1.4 represents the causal structure of this extended production function. Throughout, we use our parametrization with binary outcomes at all variables $i \in N^*$ and $p(x_i = 1 \mid x_{R(i)}) = \beta_i + \sum_{j \in R(i)} \beta_{ji} x_j$. The telemarketers either conduct cold-calls or not, $\beta_1 \in \{0, 1\}$; cold-calls have a negative effect on reputation, $\beta_{13} < 0$; consumer information has a positive effect on reputation, $\beta_{23} > 0$.[11] All formal proofs of this subsection are in Appendix 1.8.4.

Imagine that the marketer neither takes into account the telemarketers' operation nor the firm's reputation so that her subjective model is given by $\mathcal{R}$ on the upper-left of Figure 1.5. When choosing effort, she only considers the effect through consumer information. Does this misspecification change incentives? The answer is negative. We can show – using the results from the next subsection – that the agent's subjective beliefs about the production function are correct, so that $p_{\mathcal{R}}(y_H \mid a; \alpha) = p(y_H \mid a)$ for all $a \in \{0, 1\}$ and $\alpha \in [0, 1]$. Thus, given her role in the principal's project (as captured by $\mathcal{R}^*$), the subjective model $\mathcal{R}$ is rich enough to produce correct predictions. The agent may ignore important parts of the project and still act as if she were fully rational. The optimal contract is then the same as in the canonical model.

Importantly, telemarketing still matters for the principal since the probability distribution over sales depends on whether cold-calls are made or not. It is just not essential for the agent to know whether cold-calls take place. Her estimate of the production function implicitly takes into account the deterministic activity of the telemarketers.

Is there any simplification that would make the agent overestimate the effectiveness of her effort? Again, the answer is negative. If the agent does not take node 2 into account, she believes that her action has no consequences for the output. It would then be impossible

[11] Here we introduce a link between consumer information and reputation, and violate our full support assumption by assuming $p(x_1 = 1) \in \{0, 1\}$. The latter implies that in objective model $\mathcal{R}^*$ we could drop node 1 and factor the value $p(x_1 = 1)$ into the other conditional probabilities. If $p(x_1 = 1) \in (0, 1)$, node 1 would be a confounding factor and the behavioral rationality result in the example would no longer hold. In terms of interpretation, this assumption just means telemarketing either takes place or not.

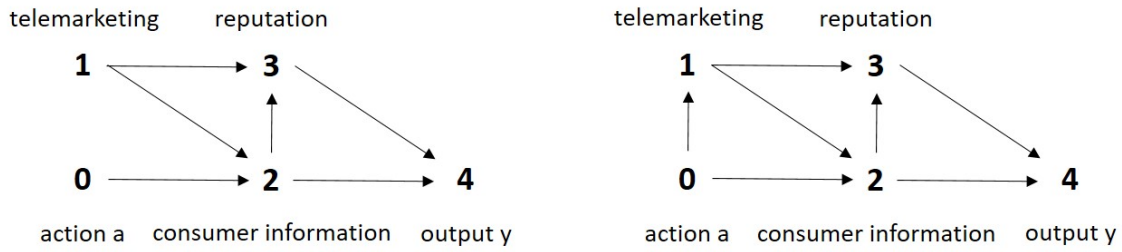

Figure 1.4: Objective model $\mathcal{R}^*$ (left) when the agent works as ordinary marketer, and objective model $\mathcal{R}^{**}$ (right) when the agent works as "head of marketing."

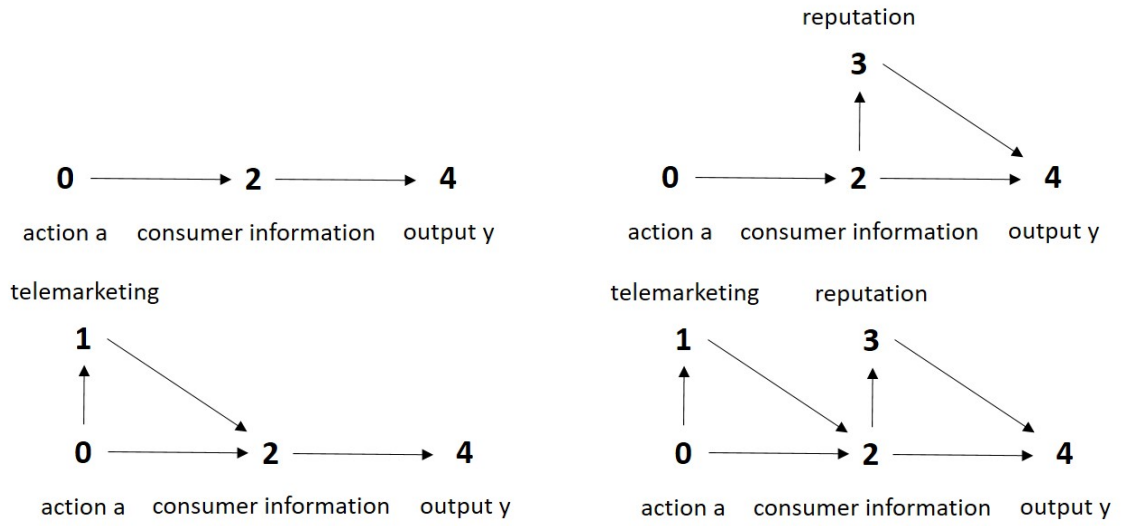

Figure 1.5: Subjective models $\mathcal{R}$ (upper-left), $\mathcal{R}_1$ (upper-right), $\mathcal{R}_2$ (lower-left), and $\mathcal{R}_3$ (lower-right).

to implement high effort. If only node 1 or only node 3 were omitted from her subjective model, the agent would again have correct beliefs about the production function. Thus, there is no scope for control optimism when the agent works as ordinary marketer.

Next, we alter the agent's job by promoting her to "head of marketing." Her action now influences the telemarketers' effort, for example, by motivating or inspiring the telemarketers. Instead of $p(x_1 = 1) = \beta_1$, we now have $p(x_1 = 1 \mid a) = \beta_1 + \beta_{01}a$. To keep things as close as possible to the previous case, we assume $\beta_1 = 0$ and $\beta_{01} = 1$.[12] Hence, the agent needs to act in order to get the telemarketers going. The objective model of the extended production function is given by $\mathcal{R}^{**}$ on the right of Figure 1.4. How does a misspecification in the agent's subjective model now affect equilibrium beliefs and incentives in this environment?

Let us first assume that the agent has the same subjective model $\mathcal{R}$ as before (on the upper-left of Figure 1.5). She neglects both the telemarketers' activity and the firm's

[12]Formally, we assume $\beta_1 = \varepsilon_1$ and $\beta_{01} = 1 - \varepsilon_2$ where $\varepsilon_1 < \varepsilon_2$, and consider the limit beliefs as $\varepsilon_1 \to 0$ and $\varepsilon_2 \to 0$. We show in the proofs for this subsection that our results do not depend on this assumption.

reputation. This is not realistic since as "head of marketing" the agent should be aware of her subordinates' basic activities; so we will relax this assumption below. The misspecification now affects incentives. Under the objective model $\mathcal{R}^{**}$ the IC that implements $\alpha = 1$ would be

$$[(\beta_{02} + \beta_{01}\beta_{12})(\beta_{24} + \beta_{23}\beta_{34}) + \beta_{01}\beta_{13}\beta_{34}](u(w(y_H)) - u(w(y_L))) \geq c. \tag{1.18}$$

The squared brackets contain the different channels through which effort affects output. The partial negative effect of effort on output through cold-calls and reputation is captured in the term $\beta_{01}\beta_{13}\beta_{34}$; it is negative since $\beta_{13} < 0$. Under the subjective model $\mathcal{R}$ the IC becomes

$$(\beta_{02} + \beta_{01}\beta_{12})(\beta_{24} + \beta_{23}\beta_{34})(u(w(y_H)) - u(w(y_L))) \geq c. \tag{1.19}$$

Here the partial negative effect is missing so that the IC is relaxed. Note that through the estimate of the link between the agent's action and consumer information, the agent implicitly takes into account her positive influence on the telemarketers' effort, which in turn positively affects consumer information (see the term $\beta_{01}\beta_{12}$). Therefore, by being promoted to a job where the agent also influences telemarketing, she overestimates her productivity. The principal benefits from this since the misspecification reduces the need to provide effort incentives.

Assume now that the agent takes the telemarketers' action into account, but still omits reputation in her model. Therefore, her subjective model is given by $\mathcal{R}_2$ on the lower-left of Figure 1.5. Does this inclusion correct, at least partly, the agent's beliefs? It turns out that this is not the case. The models $\mathcal{R}$ and $\mathcal{R}_2$ produce the same beliefs about the effectiveness of effort, i.e., $p_{\mathcal{R}}(y_H \mid a; \alpha) = p_{\mathcal{R}_2}(y_H \mid a; \alpha)$ for all $a \in \{0, 1\}$ and $\alpha \in [0, 1]$. Including more variables does not necessarily make the agent more rational. This also holds for the models $\mathcal{R}_1$ and $\mathcal{R}_3$ in Figure 1.5. Note that $\mathcal{R}_3$ is almost equal to the objective model $\mathcal{R}^{**}$, only the link between telemarketing and reputation is missing. Yet, all subjective models in this figure produce the same beliefs. Thus, a small misspecification in the agent's subjective model can render several important variables as inessential for estimating the production function.

Proposition 1.5 (Scope for Control Optimism)**.** *Consider the job examples of this subsection.*

(a) If the agent works as ordinary marketer (objective model $\mathcal{R}^*$), the misspecification in $\mathcal{R}$ has no effect on the IC and the optimal equilibrium contract is the same as in the canonical model. There is no simplification that generates control optimism.

(b) If the agent works as "head of marketing" (objective model $\mathcal{R}^{**}$), the misspecification in $\mathcal{R}$ generates control optimism and relaxes the IC; the subjective models $\mathcal{R}$, $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ generate the same beliefs about the production function.

Proposition 1.5 illustrates how the agent's job may matter for optimal incentives. The two jobs with objective models $\mathcal{R}^*$ and $\mathcal{R}^{**}$ may give rise to the same production function $p(y \mid a)$,[13] so that incentives would be identical under rational expectations. However, effort motivation is larger under a job with the objective model $\mathcal{R}^{**}$ when the agent's subjective model is simplified in a way that benefits the principal. The crucial difference between the jobs are the sets of channels through which the action affects the output. In the next subsection, we will formally define these channels.

Part (a) and (b) of Proposition 1.5 combined demonstrate that an agent's degree of control optimism may be determined by the nature of her job. In the example, the agent with misspecified model $\mathcal{R}$ was behaviorally rational in her job as ordinary marketer, but overestimated the importance of her effort after being promoted to "head of marketing" where she influences the actions of others. Thus, in our framework, the agent's control optimism is not caused by certain features of her personality, but it is a consequence of her environment when her subjective model does not capture all empirical regularities of this environment.

1.5.2 A General Result on Behavioral Rationality

To obtain a general result on behavioral rationality, we assume that the objective model $\mathcal{R}^*$ is perfect, and that the agent's subjective model $\mathcal{R}$ is a simplification. $\mathcal{R}$ will then be perfect. No v-structure emerges if we take out nodes from a perfect $\mathcal{R}^*$ and all links attached to them. The assumptions on $\mathcal{R}^*$ and $\mathcal{R}$ are not overly restrictive: Any probability distribution $p(x_{N^*})$ factorizes according to some perfect DAG $\mathcal{R}^*$. The assumption

[13] Specifically, when we denote parameters for the job with objective model $\mathcal{R}^*$ ($\mathcal{R}^{**}$) with "$*$" ("$**$") we only have to select parameters so that $\beta^*_{02}(\beta^*_{24} + \beta^*_{23}\beta^*_{34}) = (\beta^{**}_{02} + \beta^{**}_{01}\beta^{**}_{12})(\beta^{**}_{24} + \beta^{**}_{23}\beta^{**}_{34}) + \beta^{**}_{01}\beta^{**}_{13}\beta^{**}_{34}$.

on $\mathcal{R}$ is satisfied by almost all subjective models in this paper. All formal proofs for this subsection are in Appendix 1.8.5.

In the following, we characterize for any perfect $\mathcal{R}^*$ the subset of nodes the agent needs to have in her subjective model $\mathcal{R}$ so that she acts as if she had fully rational beliefs about the production function. We use the following definitions and results from the Bayesian network literature. Consider any DAG $\mathcal{R} = (N, R)$. Its skeleton $(N, \tilde{R})$ is obtained by making the DAG undirected. We have $i\tilde{R}j$ if and only if iRj or jRi.

Definition 1.5. *Two DAGs $\mathcal{R}$ and $\mathcal{G}$ are equivalent if $p_{\mathcal{R}}(x_{N^*}) \equiv p_{\mathcal{G}}(x_{N^*})$ for every $p \in \Delta(X_{N^*})$.*

Proposition 1.6 (Verma and Pearl, 1991)**.** *Two DAGs $\mathcal{R}$ and $\mathcal{G}$ are equivalent if and only if they have the same skeleton and v-structure.*

Two different models produce the same beliefs if they share the same skeleton and the same set of v-colliders. A subset of nodes $M \subset N$ is called ancestral in $\mathcal{R}$ if for all nodes $i \in M$ we have $R(i) \subset M$. A path τ of length d from node i to node j is a sequence of nodes $\tau_0, \tau_1, ..., \tau_d$ so that $\tau_0 = i$, $\tau_d = j$, and $\tau_{h-1}\tilde{R}\tau_h$ for all $h \in \{1, ..., d\}$. The length of the shortest path between i and j is called the distance between these nodes and denoted by $d(i, j)$. A path of length d is active if there is no $h \in \{1, ..., d-1\}$ so that $\tau_{h-1}R\tau_h$ and $\tau_{h+1}R\tau_h$.

Define by $\mathcal{E}$ the set of DAGs in the equivalence class of $\mathcal{R}^*$ in which the action node 0 is ancestral (nothing influences the agent's action). In each of these DAGs, all active paths between the action node 0 and any node i point towards i. Thus, the assumption that node 0 is ancestral pins down the direction of many links in a perfect DAG. We call such links "fundamental links." There is a close connection between fundamental links and the set of nodes that can be removed while maintaining behavioral rationality.

Definition 1.6. *Consider two nodes $i, j \in N^*$. If iGj for all $\mathcal{G} = (G, N^*) \in \mathcal{E}$, then the link iGj is called fundamental link and denoted by iEj.*

An intuition for fundamental links is that they capture empirically relevant directions of causality (given agreement on the ancestral node). Specifically, they describe how the agent's action impacts on other variables. Consider $\mathcal{R}^*$ from Figure **??**. Since the action node is ancestral, the links pointing from node 0 to other nodes are fundamental ($0R^*1$, $0R^*2$, and $0R^*3$). Thus, the two links pointing into the output node ($1R^*4$ and $3R^*4$) also must be fundamental. If we would turn around one of them, we would create a v-collider

since there is no link between node 0 and node 4. The remaining links $1R^*2$, $1R^*3$, and $2R^*3$ are not fundamental. We can state a result that characterizes all fundamental links in any perfect DAG; see Appendix 1.8.5. For now, we go a step further and consider sequences of fundamental links.

Definition 1.7. *Let τ be an active path in $\mathcal{R}^*$. Then τ is a fundamental active path if all the links between neighboring nodes in τ are fundamental.*

Fundamental active paths are what we so far called "channels." Consider again $\mathcal{R}^*$ from Figure **??**. The path $\tau = \{0, 1, 4\}$ is a fundamental active path since both links $0R^*1$ and $1R^*4$ are fundamental. In contrast, the active path $\tau' = \{0, 2, 3, 4\}$ is not fundamental since the link $2R^*3$ is not fundamental. We define the set of nodes that are part of at least one fundamental active path between the action and the output by

$$H^*(\mathcal{R}^*) := \{i \in N^* \mid i \text{ is part of a fundamental active path between 0 and } n \text{ in } \mathcal{R}^*\}.$$

It turns out that the nodes in $H^*(\mathcal{R}^*)$ are exactly those nodes the agent needs to have in her subjective model in order to be behaviorally rational, provided that her subjective model is a simplification. We can prove this by finding a DAG $\mathcal{G}$ that is equivalent to $\mathcal{R}^*$ and in which there are no links pointing from nodes in $N^* \setminus H^*(\mathcal{R}^*)$ to nodes in $H^*(\mathcal{R}^*)$. In this DAG, the nodes that are not in $H^*(\mathcal{R}^*)$ have no influence on the output, so the agent can safely ignore them. By Proposition 1.6, the agent correctly anticipates the production function if $H^*(\mathcal{R}^*) \subseteq N$.

Proposition 1.7 (Behavioral Rationality)**.** *Let $\mathcal{R}^*$ be a perfect DAG and let the agent's subjective DAG $\mathcal{R}$ be a simplification. The agent is behaviorally rational if and only if $\mathcal{R}$ contains all nodes from $H^*(\mathcal{R}^*)$.*

Proposition 1.7 implies that the agent does not necessarily have to take into account all variables of her (potentially) complex environment in order to be behaviorally rational. In particular, this holds independent of the parametrization of the extended production function. For example, when $p(x_1, ..., x_4 \mid a)$ factorizes according to $\mathcal{R}^*$ in Figure **??**, the agent can ignore node 2 and still would behave as in the contracting model with common priors. The intuition is that when $H^*(\mathcal{R}^*) \subseteq N$, then the information captured through the variables in $H^*(\mathcal{R}^*)$ already includes the probabilistic information from variables outside $H^*(\mathcal{R}^*)$. Conversely, if the agent's subjective model does not include all variables from $H^*(\mathcal{R}^*)$, she is not behaviorally rational. In this case, we can find a parametrization of $p(x_1, ..., x_n \mid a)$ such that the incentive compatibility constraint is affected by the

simplification in the agent's subjective model $\mathcal{R}$.

Next, Proposition 1.7 also shows that different misspecifications can have the same effect on incentives. Consider the two models $\mathcal{R}_1$ and $\mathcal{R}_2$ from the job example in Figure 1.5. The set of nodes on fundamental active paths is the same for these two models, $H^*(\mathcal{R}_1) = H^*(\mathcal{R}_2) = \{0, 2, 4\}$. This implies that the agent's beliefs under these models are identical. Thus, it does not matter for the equilibrium contract whether the agent ignores node 1, node 3, or both nodes. Therefore, the ignorance about one channel of causality may render another variable unimportant. A further interpretation is that two agents with different subjective models may have the same beliefs about the production function. We capture this result in a general statement. Consider a DAG $\mathcal{R} = (N, R)$ and a subset $\tilde{N} \subset N$. Denote by $\mathcal{R}^{[\tilde{N}]} = (\tilde{N}, \tilde{R})$ with $\tilde{R} = (\tilde{N} \times \tilde{N}) \cap R$ the DAG $\mathcal{R}$ restricted on $\tilde{N}$.

Corollary 1.3. *Let $\mathcal{R}_1 = (N_1, R_1)$ and $\mathcal{R}_2 = (N_2, R_2)$ be two perfect DAGs. Suppose there exists a DAG $\mathcal{R}_3$ so that $\mathcal{R}_3^{[N_1]} = \mathcal{R}_1$ and $\mathcal{R}_3^{[N_2]} = \mathcal{R}_2$. If $H^*(\mathcal{R}_1) = H^*(\mathcal{R}_2)$, then we have that $p_{\mathcal{R}_1}(y \mid a; q) = p_{\mathcal{R}_2}(y \mid a; q)$ for all $a \in A$ and $q \in \Delta(A)$.*

Finally, note that one can make any imperfect DAG perfect by adding links between nodes that create v-colliders. If $p(x_{N^*})$ is consistent with $\mathcal{R}^*$, it is consistent with any DAG that adds links to $\mathcal{R}^*$. One can exploit this to partially extend Proposition 1.7 to imperfect objective models.

1.6 Comparative Statics

One advantage of our approach to contracting with boundedly rational agents is that beliefs are derived endogenously from the true production process. This allows us to analyze how the optimal equilibrium contract varies in the parameters of the environment. As an example, we briefly revisit the trade-off between risk and incentives. This comparative static that has been discussed extensively in the contracting literature.[14]

In the canonical contracting model, the trade-off works as follows. A risk-averse agent demands a risk premium for accepting a wage schedule with uncertain wage payments. Thus, an increase in risk drives up the costs of providing incentives. Consequently, the provision of effort incentives should decrease in the riskiness of the environment. However,

[14] In an earlier version of the paper, we also discussed the trade-off between team size and incentives (it is available upon request from the authors).

empirically this relationship does not hold in general (e.g., Prendergast, 2002). Field evidence on the relationship between risk an incentives for CEO compensation is mixed, and for other domains, such as franchising, a positive relationship can be observed. In contrast, a negative relationship is obtained in lab experiments where subjects know the true production function (Corgnet and Hernán-González, 2019).

We can use our marketer example to show how the relationship between risk and incentives may become positive when the agent has a simplified model of the project; see Appendix 1.8.6 for details. We consider a mean-preserving spread in $p(y \mid a)$, so that under the objective model $\mathcal{R}^*$ the provision of incentives becomes more costly when there is more risk. However, if the agent's subjective model is misspecified, there can be an additional effect of risk on incentives: The agent may perceive the riskier environment as one in which her action is more important for the output. This relaxes the incentive compatibility constraint. If this effect is sufficiently strong relative to the risk premium effect, there can be a positive relationship between risk and incentives.

1.7 Conclusion

In this paper, we applied Spiegler's (2016) Bayesian network framework to analyze optimal contracting in a principal-agent setting where the agent forms beliefs about the production function based on a misspecified model of the principal's project. The objective causal model may be very complex, and may contain empirical regularities that the agent does not consider due to cognitive limitations or because they are never brought to her attention.

The optimal contract exhibits the following features. First, it does not exploit the agent if her subjective model takes into account the correlation between variables in her model that have a joint influence on a third variable (in which case it is "perfect"). Second, the principal may nevertheless benefit from a misspecification in the agent's perfect subjective model if it makes the agent control optimistic so that the incentive compatibility constraint is relaxed. Third, if the agent's subjective model is perfect, the agent cannot infer from the shape of incentives that her beliefs are biased. Fourth, when the agent correctly anticipates the joint distribution of contractible variables, the optimal contract conditions on an additional variable only if it is informative about the action according to the agent's model. Fifth, the optimal contract is identical to the rational benchmark if the agent is behaviorally rational. We characterize when this is the case, and apply this finding to show how the scope for control optimism may depend on the agent's job. For example,

a front-line worker may not fully understand the workings of the organization around her, but still act as if she were fully rational. In contrast, a high-ranking manager, who affects the output by influencing the behavior of many subordinates, overestimates her own productivity if she does not take into account the challenges that her subordinates face in their routines.

We focused on a simple contracting framework so that we can identify precisely how misspecifications in the agent's model affect incentive contracts. Future research can extend the framework by considering team incentives, relational contracts, and delegation. The Bayesian network approach offers a very disciplined tool to study the effects of bounded rationality on organizations, and we think that our results are useful in this respect.

References

Sarah Auster. Asymmetric awareness and moral hazard. *Games and Economic Behavior*, 82:503–521, 2013.

Lucian Arye Bebchuk and Jesse Fried. Pay without performance: The unfulfilled promise of executive compensation. *Harvard University Press, Cambridge*, 2004.

Roland Bénabou. Groupthink: Collective delusions in organizations and markets. *Review of Economic Studies*, 80:429–462, 2013.

Roland Bénabou and Jean Tirole. Self-confidence and personal motivation. *Quarterly Journal of Economics*, 117:871–915, 2002.

Marianne Bertrand and Sendhil Mullainathan. Are CEOs rewarded for luck? the ones without principals are. *Quarterly Journal of Economics*, 116:901–932, 2001.

Nicolas Bloom, Benn Eifert, Aprajit Mahjan, David McKenzie, and John Roberts. Does management matter? evidence from India. *Quarterly Journal of Economics*, 128:1–51, 2013.

Markus Brunnermeier and Jonathan Parker. Optimal expectations. *American Economic Review*, 95:1092–1118, 2005.

Pierre Chaigneau, Alex Edmans, and Daniel Gottlieb. The informativeness principle wihtout the first-order approach. *Games and Economic Behavior*, 113:743–755, 2019.

Brice Corgnet and Roberto Hernán-González. Revisiting the trade-off between risk and incentives: The shocking effect of random shocks? *Management Science*, 65:1096–1114, 2019.

Enrique De la Rosa. Overconfidence and moral hazard. *Games and Economic Behavior*, 73:429–451, 2011.

Marci DeCaro, Robin Thomas, Niel Albert, and Sian Beilock. Choking under pressure: Multiple routes to skill failure. *Journal for Experimental Psychology: General*, 140: 390–406, 2011.

Eddie Dekel, Barton Lipman, and Aldo Rustichini. Standard state-space models preclude unawareness. *Econometrica*, 66:159–173, 1998.

Kfir Eliaz and Ran Spiegler. A model of competing narratives. *American Economic Review*, 110(3786-3816):3786–3816, 2020.

Kfir Eliaz, Ran Spiegler, and Heidi C. Thysen. Strategic interpretations. *Journal of Economic Theory*, 192:105192, 2021.

Hanming Fang and Giuseppe Mocarini. Morale hazard. *Journal of Monetary Economics*, 52:749–777, 2005.

Emel Filiz-Ozbay. Incorporating unawareness into contract theory. *Games and Economic Behavior*, 76:181–194, 2012.

Simon Gervais and Itay Goldstein. The positive effects of biased self-perceptions in firms. *Review of Finance*, 11:453–496, 2007.

Sanford Grossman and Oliver Hart. An analysis of the principal-agent problem. *Econometrica*, 51:7–45, 1983.

Rema Hanna, Sendhil Mullainathan, and Joshua Schwartzstein. Learning through noticing: Theory and evidence from a field experiment. *Quarterly Journal of Economics*, 129:1311–1353, 2014.

Aviad Heifetz, Martin Meier, and Burkhard Schipper. Interactive unawareness. *Journal of Economic Theory*, 130:78–94, 2006.

Aviad Heifetz, Martin Meier, and Burkhard Schipper. Unawareness, beliefs, and speculative trade. *Games and Economic Behavior*, 77:100–121, 2013.

Bengt Holmström. Moral hazard and observability. *Bell Journal of Economics*, 10:74–91, 1979.

Giovanni Immordino, Anna Maria Menchini, and Maria Grazia Romano. Contracts with wishful thinkers. *Journal of Economics and Management Strategy*, 24:863–886, 2015.

Botond Kőszegi. Ego utility, overconfidence, and task choice. *Journal of European Economic Association*, 4:673–707, 2006.

Botond Kőszegi. Behavioral contract theory. *Journal of Economic Literature*, 52:1075–1118, 2014.

Timo Koski and John Noble. *Bayesian Networks - An Introduction*. John Wiley & Sons, 2009.

Jeanine Miklós-Thal and Juanjuan Zhang. (de)marketing to manage consumer quality inferences. *Journal of Marketign Research*, 50:55–69, 2013.

Sherwin Nuland. *The doctors' plague: Germs, Childbed Fever, and the strange story of Ignac Semmelweis*. W. W. Norton Company, 2004.

Judea Pearl. *Causality: Models, Reasoning, and Inferences*. Cambridge University Press, 2009.

Micheal Porter, Jay Lorsch, and Nitin Nohria. Seven surprises for new CEOs. *Harvard Business Review*, 82:62–72, 2004.

Canice Prendergast. The tenuous trade-off between risk and incentives. *Journal of Political Economy*, 110:1071–1102, 2002.

Luís Santo-Pinto. Positive self-image and incentives in organisations. *Economic Journal*, 118:1315–1332, 2008.

Anja Sautmann. Self-confidence in a principal-agent relationship. 2007.

Anja Sautmann. Contracts for agents with biased beliefs: Some theory and an experiment. *American Economic Journal: Microeconomics*, 5:124–156, 2013.

Pablo Schenone. Causality: A decision theoretic framework. *Working Paper, California Institute of Technology*, 2020.

Herbert Simon. Administrative behavior. *Macmillan, London*, 1947.

Herbert Simon. A behavioral model of rational choice. *Quarterly Journal of Economics*, 69:99–118, 1955.

Ran Spiegler. Bayesian networks and boundedly rational expectations. *Quarterly Journal of Economics*, 131:1243–1290, 2016.

Ran Spiegler. Data monkeys: A procedural model of extrapolation from partial statistics. *Review of Economic Studies*, 84:1818–1841, 2017.

Ran Spiegler. Can agents with causal misperceptions be systematically fooled? *Journal of the European Economic Association*, 18:583–617, 2020.

Johannes Spinnewijn. Insurance and perceptions: How to screen optimists and pessimists. *Economic Journal*, 123:606–633, 2013.

Johannes Spinnewijn. Unemployed but optimistic: Optimal insurance design with biased beliefs. *Journal of the European Economic Association*, 13:130–167, 2015.

Eric Van den Steen. Organizational beliefs and managerial vision. *Journal of Law, Economics, and Organization*, 21:256–283, 2005.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. *Uncertainty in Artifical Intelligence*, 6:255–268, 1991.

Ernest-Ludwig von Thadden and Xiaojian Zhao. Incentives for unaware agents. *Review of Economic Studies*, 79:1151–1174, 2012.

Ernest-Ludwig von Thadden and Xiaojian Zhao. Multitask agency with unawareness. *Theory and Decision*, 77:197–222, 2014.

1.8 Appendix

1.8.1 Existence of a Personal Equilibrium

We first show that the agent's subjective beliefs $p_{\mathcal{R}}(y \mid a; q)$ are well-defined at any pair a, y and for any pure or mixed strategy $q \in \Delta(A)$. Define $p_{\mathcal{R}}(y \mid a; q) = \lim_{k\to\infty} p_{\mathcal{R}}(y \mid a; q^k)$ for a sequence $q^1, q^2, \ldots$ of fully mixed strategies with the property that $q^k \to q$ as $k \to \infty$. Let $\hat{q}^1, \hat{q}^2, \ldots$ be any alternative sequence of fully mixed strategies with $\hat{q}^k \to q$ as $k \to \infty$. This implies that $s^k \to 0$ as $k \to \infty$, where $s^k = q^k - \hat{q}^k$ for $k \in \mathbb{N}$. Since $p_{\mathcal{R}}(y \mid a; \cdot)$ is continuous, we have $p_{\mathcal{R}}(y \mid a; \hat{q}^k) - p_{\mathcal{R}}(y \mid a; q^k) \to 0$ as $k \to \infty$, which proves the statement. Next, we show that a personal equilibrium exists at any admissible $\mathcal{R}$ and $w \in W$. Note that $\Delta(A)$ is non-empty, compact, and convex. Define the best-response correspondence $BR : \Delta(A) \to \Delta(A)$ by

$$BR(q) = \arg\max_{\tilde{q}\in\Delta(A)} \sum_{a'\in A} \sum_{y\in Y} \tilde{q}(a')[p_{\mathcal{R}}(y \mid a'; q)u(w(y)) - c(a')]. \tag{A.1}$$

For every $q \in \Delta(A)$ we have that $BR(q)$ is non-empty and convex. The latter statement follows since any convex combination of pure actions that are optimal for the agent is an element of $BR(q)$. Since $p_{\mathcal{R}}(y \mid a'; q)$ is continuous in q, we also must have that $\sum_{a'\in A} \sum_{y\in Y} \tilde{q}(a')[p_{\mathcal{R}}(y \mid a'; q)u(w(y)) - c(a')]$ is continuous in q. Hence, $BR(q)$ is upper hemi-continuous. The existence of a personal equilibrium then follows from Kakutani's theorem.

1.8.2 Omitted Proofs from Section 3

Proof of Proposition 1.3. Statement (a) is proven in the main text. We prove statement (b). We assume w.l.o.g. that $A = \{0, 1\}$ and $Y = \{y_L, y_H\}$, with the usual interpretation. Since the principal strictly prefers (w^*, q^*) to the optimal contract under the objective model $\mathcal{R}^*$, and the agent correctly anticipates the equilibrium distribution over output, the equilibrium action must be $a^* = 1$ and $w^*(1) > w^*(0)$. We show that from the agent's perspective the principal cannot gain by implementing $a = 0$. Denote by $\bar{w}$ the fixed wage that implements $a = 0$ at lowest costs to the principal under the objective model. The agent anticipates that a fixed wage of $\bar{w}$ would optimally implement $a = 0$. Since Y is

binary we must have $p(y_H \mid a = 0) > p_{\mathcal{R}}(y_H \mid a = 0; a^*)$. Thus, we get

$$\begin{aligned}\sum_{y \in Y} p_{\mathcal{R}}(y \mid a = 1; a^*)(y - w^*(y)) &= \sum_{y \in Y} p(y \mid a = 1)(y - w^*(y)) \\ &> \sum_{y \in Y} p(y \mid a = 0)(y - \bar{w}) \\ &> \sum_{y \in Y} p_{\mathcal{R}}(y \mid a = 0; a^*)(y - \bar{w}),\end{aligned}$$

where the first inequality follows from the fact that the principal strictly prefers (w^*, q^*) to the optimal contract under model $\mathcal{R}^*$. This completes the proof of statement (b). □

1.8.3 Omitted Proofs from Section 4

Proof of Proposition 1.4. We first prove statement (b). Suppose the principal wishes to implement q. Since the agent is risk-averse with unlimited liability and her action set A is finite, we can use the arguments in Grossman and Hart (1983) to show that the Kuhn-Tucker theorem yields necessary and sufficient conditions for an optimum. The optimal incentive scheme is therefore characterized by the first-order condition

$$\frac{1}{u'(w(y,z))} = \frac{p_{\mathcal{R}}(y,z;q)}{p(y,z)} \left[\mu + \sum_{a' \in A} \lambda_{a'} \frac{p_{\mathcal{R}}(y,z \mid a;q) - p_{\mathcal{R}}(y,z \mid a';q)}{p_{\mathcal{R}}(y,z;q)} \right] \tag{A.2}$$

for any a in the support of q. By assumption, we have $p_{\mathcal{R}}(y,z;q) = p(y,z)$. We can rewrite $p_{\mathcal{R}}(y,z \mid a;q)$ as

$$p_{\mathcal{R}}(y,z \mid a;q) = p_{\mathcal{R}}(y \mid a;q) p_{\mathcal{R}}(z \mid y,a;q) = p_{\mathcal{R}}(y \mid a;q) p_{\mathcal{R}}(z \mid y;q), \tag{A.3}$$

where the last equality follows from the assumption $p_{\mathcal{R}}(z \mid y,a;q) = p_{\mathcal{R}}(z \mid y;q)$ for all triples a, y, z. Similarly, we can write $p_{\mathcal{R}}(y,z;q) = p_{\mathcal{R}}(y;q) p_{\mathcal{R}}(z \mid y;q)$. Hence, we get

$$p_{\mathcal{R}}(y,z \mid a;q) - p_{\mathcal{R}}(y,z \mid a';q) = \frac{p_{\mathcal{R}}(y,z;q)}{p_{\mathcal{R}}(y;q)} [p_{\mathcal{R}}(y \mid a;q) - p_{\mathcal{R}}(y \mid a';q)]. \tag{A.4}$$

The first-order condition in (A.2) therefore simplifies to

$$\frac{1}{u'(w(y,z))} = \mu + \sum_{a' \in A} \lambda_{a'} \frac{p_{\mathcal{R}}(y \mid a;q) - p_{\mathcal{R}}(y \mid a';q)}{p_{\mathcal{R}}(y;q)}. \tag{A.5}$$

Since the right-hand side of this first-order equation is independent of z, the optimal incentive scheme does not condition on z, which completes the proof. Next, we prove statement (a). Risk-aversion and unlimited liability imply that the optimal incentive

scheme that implements $a = 1$ is characterized by the first-order condition

$$\frac{1}{u'(w(y,z))} = \frac{p_{\mathcal{R}}(y,z \mid a=1;\alpha=1)}{p(y,z \mid a=1)} \left[\mu + \lambda \left(1 - \frac{p_{\mathcal{R}}(y,z \mid a=0;\alpha=1)}{p_{\mathcal{R}}(y,z \mid a=1;\alpha=1)} \right) \right], \quad \text{(A.6)}$$

where μ, λ are strictly positive constants. As above, we can write $p_{\mathcal{R}}(y,z \mid a=1;\alpha=1) = p(y,z \mid a=1)$, so that this first-order condition simplifies to

$$\frac{1}{u'(w(y,z))} = \mu + \lambda \left(1 - \frac{p_{\mathcal{R}}(y,z \mid a=0;\alpha=1)}{p_{\mathcal{R}}(y,z \mid a=1;\alpha=1)} \right). \quad \text{(A.7)}$$

Statement (a) then directly follows from this equation. □

1.8.4 Omitted Proofs from Subsection 5.1

We first derive the IC under the objective model $\mathcal{R}^{**}$. The probabilities of high output after high and low effort, respectively, are given by

$$\begin{aligned} p(y_H \mid a=1) &= \beta_4 + [\beta_2 + \beta_{02} + (\beta_1 + \beta_{01})\beta_{12}]\beta_{24} \\ &\quad + [\beta_3 + (\beta_1 + \beta_{01})\beta_{13} + (\beta_2 + \beta_{02} + (\beta_1 + \beta_{01})\beta_{12})\beta_{23}]\beta_{34}, \quad \text{(A.8)} \\ p(y_H \mid a=0) &= \beta_4 + [\beta_2 + \beta_1\beta_{12}]\beta_{24} + [\beta_3 + \beta_1\beta_{13} + (\beta_2 + \beta_1\beta_{12})\beta_{23}]\beta_{34}, \quad \text{(A.9)} \end{aligned}$$

so that the effect of effort on the probability of high output equals

$$p(y_H \mid a=1) - p(y_H \mid a=0) = (\beta_{02} + \beta_{01}\beta_{12})(\beta_{24} + \beta_{23}\beta_{34}) + \beta_{01}\beta_{13}\beta_{34}. \quad \text{(A.10)}$$

Next, we drive the IC under the subjective model $\mathcal{R}$ when the equilibrium action is $\alpha \in [0,1]$. We calculate

$$\begin{aligned} p(x_1 = 1 \mid x_2 = 1) &= \frac{\alpha(\beta_1 + \beta_{01})(\beta_2 + \beta_{02} + \beta_{12}) + (1-\alpha)\beta_1(\beta_2 + \beta_{12})}{\beta_2 + \beta_1\beta_{12} + \alpha(\beta_{02} + \beta_{01}\beta_{12})}, \quad \text{(A.11)} \\ p(x_1 = 1 \mid x_2 = 0) &= \frac{\alpha(\beta_1 + \beta_{01})(1 - \beta_2 - \beta_{02} - \beta_{12}) + (1-\alpha)\beta_1(1 - \beta_2 - \beta_{12})}{1 - \beta_2 - \beta_1\beta_{12} - \alpha(\beta_{02} + \beta_{01}\beta_{12})}, \quad \text{(A.12)} \end{aligned}$$

and

$$\begin{aligned} p(x_3 = 1 \mid x_2 = 1) &= \beta_3 + p(x_1 = 1 \mid x_2 = 1)\beta_{13} + \beta_{23}, \quad \text{(A.13)} \\ p(x_3 = 1 \mid x_2 = 0) &= \beta_3 + p(x_1 = 1 \mid x_2 = 0)\beta_{13}. \quad \text{(A.14)} \end{aligned}$$

The agent's belief about the probability of high output after $x_2 = 1$ and $x_2 = 0$, respectively, is therefore given by

$$p(y_H \mid x_2 = 1) = \beta_4 + \beta_{24} + [\beta_3 + p(x_1 = 1 \mid x_2 = 1)\beta_{13} + \beta_{23}]\beta_{34}, \quad \text{(A.15)}$$

$$p(y_H \mid x_2 = 0) = \beta_4 + [\beta_3 + p(x_1 = 1 \mid x_2 = 0)\beta_{13}]\beta_{34}. \quad \text{(A.16)}$$

The agent correctly anticipates $p(x_2 \mid a)$. Hence, her belief about the effect of effort on the probability of high output under $\mathcal{R}$ equals

$$\begin{aligned} p_{\mathcal{R}}(y_H \mid a = 1; \alpha) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha) &= (\beta_{02} + \beta_{01}\beta_{12})(\beta_{24} + \beta_{23}\beta_{34}) + (\beta_{02} + \beta_{01}\beta_{12})\beta_{13}\beta_{34} \\ &\quad \times [p(x_1 = 1 \mid x_2 = 1) - p(x_1 = 1 \mid x_2 = 0)]. \quad \text{(A.17)} \end{aligned}$$

Recall that $\beta_{13} < 0$. By comparing (A.10) and (A.17) we get that at $\alpha = 1$ the misspecification in $\mathcal{R}$ relaxes the IC if and only if

$$\beta_{01} > \frac{\beta_{12}(\beta_1 + \beta_{01})(1 - \beta_1 - \beta_{01})(\beta_{02} + \beta_{01}\beta_{12})}{(1 - \beta_2 - \beta_{02} - \beta_{12}(\beta_1 + \beta_{01}))(\beta_2 + \beta_{02} + \beta_{12}(\beta_1 + \beta_{01}))}, \quad \text{(A.18)}$$

which implies the statement in the main text.

Proof of Proposition 1.5 . We prove the statements in (a). Since $\beta_1 \in \{0, 1\}$, we can rewrite the probability model without variable 1. The corresponding objective model $\tilde{\mathcal{R}}^*$ equals $\mathcal{R}^*$ in Figure 1.4 without node 1. We now apply Propositions 1.7 and 1.8. In model $\tilde{\mathcal{R}}^*$, node 3 is not on a fundamental active path. Hence, the agent with subjective model $\mathcal{R}$ is behaviorally rational, which yields the results. We prove the statements in (b). The first statement is shown in the text. The second statement follows from Corollary 3. Note that, in all models of Figure 1.5, the set of nodes on fundamental active paths is identical. □

1.8.5 Omitted Proofs from Subsection 5.2

To prove Proposition 1.7, we first state and prove Proposition 1.8 below.

Proposition 1.8 (Fundamental Links)**.** *Let $\mathcal{R}^*$ be a perfect DAG and consider two adjacent nodes $i, j \in N^*$. The link iR^*j is fundamental if and only if at least one of the following conditions is satisfied:*

(a) we have $d(0, i) = d(0, j) - 1$;

(b) there exists a node $k \in N^*$ such that kEi and $k \notin R^*(j)$.

This result shows that nodes that are connected by fundamental links in perfect DAGs exhibit characteristics that are easy to identify. It is not always simple to spot the nodes that are not in $H^*(\mathcal{R}^*)$. In this case, Proposition 1.8 is helpful. Consider, for example, the perfect DAG $\mathcal{R}^*$ in Figure 1.6. Condition (a) from Proposition 1.8 implies that all links which connect nodes of different distances to the action node are fundamental. The remaining links are $1R^*2$, $3R^*4$, $3R^*5$, $4R^*5$, $4R^*6$, and $5R^*6$. Condition (b) from Proposition 1.8 then implies that $4R^*6$ and $5R^*6$ are fundamental links, while the remaining links are non-fundamental. We therefore get $H^*(\mathcal{R}^*) = N^* \setminus \{3\}$.

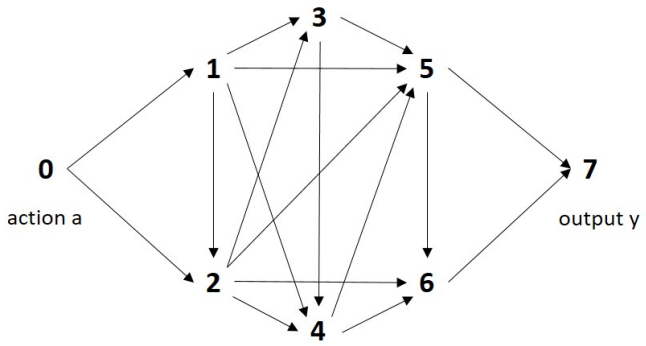

Figure 1.6: Example DAG $\mathcal{R}^*$

In order to prove Proposition 1.8, we show several intermediate results. We first note that, in a perfect DAG $\mathcal{R}^*$, the link iR^*j is fundamental if the nodes i and j differ in their distance to the action node 0.

Lemma 1.1. *Let $i, j \in N^*$ be adjacent nodes in $\mathcal{R}^*$. If $d(0,i) = d(0,j) - 1$, then iEj.*

Proof. First, suppose $d(0,i) = 0$ so that $i = 0$. Since node 0 is ancestral, we must have iGj in every DAG $\mathcal{G} \in \mathcal{E}$. Next, suppose $d(0,i) = d > 0$. Since $\mathcal{R}^*$ is perfect and node 0 is ancestral, there exists an active path of length d from node 0 to node i. Denote by k the direct ancestor of i on this path. There cannot exist a link between k and j, otherwise we would have $d(0,i) = d(0,k)$, a contradiction. Thus, we must have iGk in every DAG $\mathcal{G} \in \mathcal{E}$, otherwise we would have a v-collider at node i. □

Lemma 1.2. *Let $i, j \in N^*$ and iR^*j. If there exists a node $k \in N^*$ such that kEi and $k \notin R^*(j)$, then iEj.*

Proof. If there is a fundamental link from node k to node i, then iR^*j implies that we cannot have jR^*k. Otherwise, we would have a directed cycle. Node j and node k are therefore not adjacent. Hence, if jGi in some DAG $\mathcal{G} \in \mathcal{E}$, there would be a v-collider at i, a contradiction. □

The "if"-statement of Proposition 1.8 follows directly from Lemma 1 and Lemma 2. For the "only if"-statement we need two more results. The first one provides a condition under which a link is not fundamental.

Lemma 1.3. *Let $i, j \in N^* \setminus \{0\}$ and iR^*j. If $R^*(i) \subset R^*(j)$, then the link between i and j is not fundamental.*

Proof. Consider the DAG $\mathcal{G} = (G, N^*)$ that is identical to $\mathcal{R}^*$ except that it reverses the link between i and j. The assumption $R^*(i) \subset R^*(j)$ rules out that there are v-colliders in $\mathcal{G}$. Assume that there is a cycle in $\mathcal{G}$. Since $\mathcal{R}^*$ is acyclic, the cycle must contain jGi. Further, there must exists a node k and a link kGj which is part of the cycle. Since $\mathcal{R}^*$ is perfect, we must have $k\tilde{R}^*i$. Assume first that we have kR^*i. Then jGi implies that kGi is not part of the cycle. Thus, there must exist an active path τ of some length d so that $\tau_0 = i$ and $\tau_d = k$. But then there is a cycle consisting of the link kGi and τ. This cycle also exists in $\mathcal{R}^*$, a contradiction. Next, assume that we have iR^*k. Since $i \neq 0$ and $R^*(i) \subset R^*(j)$, there exists a node l with lR^*i and lR^*j. Since $\mathcal{R}^*$ is perfect, we also must have $l\tilde{R}^*k$. The same applies to all $l' \in R^*(i)$. Hence, starting from $\mathcal{R}^*$, we can reverse the links between i and j as well as between i and k and obtain a DAG $\mathcal{G}' \in \mathcal{E}$. □

The second result needed for the proof of the "only if"-statement of Proposition 1.8 demonstrates that for each node i in a perfect DAG $\mathcal{R}^*$ there exists a DAG $\mathcal{G} \in \mathcal{E}$ in which there is no non-fundamental link that points towards i.

Lemma 1.4. *For all nodes $i \in N^*$ there exists a DAG $\mathcal{G} \in \mathcal{E}$ in which all non-fundamental links adjacent to node i point away from i.*

Proof. Let N_d be the set of nodes that have distance $d > 0$ to the action node 0. Denote by $N_d^{[\kappa]}$, $\kappa = 1, 2, ...$, the maximal subset of nodes that (i) are at distance $d > 0$ from the action node 0, and (ii) are connected through non-fundamental links (i.e., for any two nodes $i, j \in N_d^{[\kappa]}$ there exists a path between i and j consisting of non-fundamental links). **Step 1.** We show that all nodes in a given set $N_d^{[\kappa]}$ have the same parents outside of $N_d^{[\kappa]}$. Consider two nodes $i, j \in N_d^{[\kappa]}$ that are connected through the non-fundamental link iR^*j. By definition, we have kEi for each $k \in R^*(i) \setminus N_d^{[\kappa]}$ for each $i \in N_d^{[\kappa]}$. Since $\mathcal{R}^*$ is perfect, this implies that $R^*(j) \setminus N_d^{[\kappa]} \subset R^*(i) \setminus N_d^{[\kappa]}$. Since iR^*j is non-fundamental, we also must have $R^*(i) \setminus N_d^{[\kappa]} \subset R^*(j) \setminus N_d^{[\kappa]}$ so that $R^*(i) \setminus N_d^{[\kappa]} = R^*(j) \setminus N_d^{[\kappa]}$. The result follows from the fact that, by assumption, all nodes in $N_d^{[\kappa]}$ are connected through non-fundamental links. **Step 2.** Consider two links $i \in N_d^{[\kappa]}$ and $i' \in N_d^{[\kappa']}$ with $\kappa \neq \kappa'$ that are adjacent. Assume w.l.o.g. that iR^*i'. By definition, iR^*i' is a fundamental link. Step 1 then implies that iEj' for all $j' \in N_d^{[\kappa']}$. Thus, there cannot exist nodes $j \in N_d^{[\kappa]}$ and $j' \in N_d^{[\kappa']}$ so that $j'R^*j$. Otherwise, we would have $j'Ej$ and $j'Ei$ for all $i \in N_d^{[\kappa]}$, a contradiction. Thus, there cannot exist nodes $i, j \in N_d^{[\kappa]}$ and $i', j' \in N_d^{[\kappa']}$ such that iR^*i' and $j'R^*j$. **Step 3.** Note that, since $\mathcal{R}^*$ is perfect, by Lemma 1 all links between N_d and N_{d+1} point away from the nodes in N_d. **Step 4.** We now can prove Lemma 4. Take any node $i \in N^*$ and assume w.l.o.g. that $i \in N_d^{[\kappa]}$. Consider the DAG $\mathcal{G}^{[\kappa]} = (N_d^{[\kappa]}, G^{[\kappa]})$ where $G^{[\kappa]}$ is identical to R^* restricted on $N_d^{[\kappa]}$. Since $\mathcal{R}^*$ is perfect, $\mathcal{G}^{[\kappa]}$ also must be perfect. Corollary 1 from Spiegler (2019) implies that there exists a DAG $\mathcal{Q}^{[\kappa]}$ in which node i is ancestral and that is equivalent to $\mathcal{G}^{[\kappa]}$. Choose such a $\mathcal{Q}^{[\kappa]}$ and replace $\mathcal{G}^{[\kappa]}$ in the original DAG $\mathcal{R}^*$ by $\mathcal{Q}^{[\kappa]}$. Call the resulting DAG $\mathcal{Q}^*$. Step 1 implies that there are no v-colliders in $\mathcal{Q}^*$, and Step 2 and 3 imply that there are no cycles in $\mathcal{Q}^*$, which proves the result. □

Proof of Proposition 1.8. The "if"-statement follows from Lemma 1 and Lemma 2. We prove the "only if"-statement. Consider any two adjacent nodes $i, j \in N^*$ with iR^*j and $d(0, i) = d(0, j)$. Suppose that for any node $k \in R^*(i)$ with a fundamental link kR^*i we also have $k \in R^*(j)$. By Lemma 4, we can find a DAG $\mathcal{G} \in \mathcal{E}$ in which all non-fundamental links are turned away from node i. In this DAG, we have $G(i) \subset G(j)$. From Lemma 3 it then follows that the link iR^*j is not fundamental. This completes the proof. □

Before we can prove Proposition 1.7, we need two more results. We will use the following definitions. Recall that a path τ of length d is directed if for any $h \in \{1, ..., d\}$ we have $\tau_{h-1} R \tau_h$ on this path. For any DAG, the topological ordering is a sequence of nodes such that every link is directed from an earlier to a later node in the sequence.

Lemma 1.5. *Let $M \subset N^* \setminus H^*(\mathcal{R}^*)$ be a set of nodes connected through non-fundamental links. Suppose there are two nodes $i, j \in H^*(\mathcal{R}^*)$ with non-fundamental links to nodes in M. Then i and j are adjacent.*

Proof. As in the proof of Lemma 4, let N_d be the set of nodes that have distance $d > 0$ to the action node 0. Let $E(i)$ be the set of nodes k with kEi. By Lemma 1, there is a $d > 0$ so that $i, j \in N_d$ and $M \subset N_d$. By Lemma 2, we must have $E(i) = E(j)$ since these nodes are connected through non-fundamental links. Choose any node $k \in N_{d-1}$ with $k \in H^*(\mathcal{R}^*)$ and kR^*i. By Lemma 2, we also have kR^*j. We can now choose two fundamental active paths $\tau^{[i]}, \tau^{[j]}$ from node 0 to node n so that (i) $k \in \tau^{[i]}$ and $k \in \tau^{[j]}$, (ii) $i \in \tau^{[i]}$ and $j \in \tau^{[j]}$, (iii) all nodes on $\tau^{[i]}$ and $\tau^{[j]}$ before k are identical, and (iv) there is not any node on $\tau^{[i]}$ ($\tau^{[j]}$) between k and i (k and j). Since $i, j \in H^*(\mathcal{R}^*)$ this is possible. Now define by $m_1^{[i]}$ ($m_1^{[j]}$) the last node on $\tau^{[i]}$ ($\tau^{[j]}$) before node n; by $m_2^{[i]}$ ($m_2^{[j]}$) the penultimate node on $\tau^{[i]}$ ($\tau^{[j]}$) before node n, and so forth. Since $\mathcal{R}^*$ is perfect, $m_1^{[i]}$ and $m_1^{[j]}$ must be adjacent. Since $m_1^{[i]}$ and $m_1^{[j]}$ are adjacent and $\mathcal{R}^*$ is perfect, $m_2^{[i]}$ and $m_2^{[j]}$ must be adjacent, and so forth. If nodes i and j are both the t'th node from n in $\tau^{[i]}$ ($\tau^{[j]}$), we are done. Assume that this is not the case, and that w.l.o.g. node i is the t'th node from n while node j is the t''th node from n, with $t' > t$. Then i is adjacent to $m_t^{[j]}$, and also to all nodes on $\tau^{[j]}$ between $m_t^{[j]}$ and j (including j) through non-fundamental links, otherwise there would be a contradiction to $E(i) = E(j)$. □

The next result is crucial for the proof of Proposition 1.7. It shows that all nodes that are not on a fundamental active path between action and output can be made "unimportant", in the sense that we can find a DAG in $\mathcal{E}$ in which any link between a node in $H^*(\mathcal{R}^*)$ and a node in $N^* \setminus H^*(\mathcal{R}^*)$ points towards the node in $N^* \setminus H^*(\mathcal{R}^*)$.

Lemma 1.6. *There exists a DAG $\mathcal{G}^* \in \mathcal{E}$ such that in $\mathcal{G}^*$ all links with one end in $H^*(\mathcal{R}^*)$ and the other in $N^* \setminus H^*(\mathcal{R}^*)$ point from $H^*(\mathcal{R}^*)$ to $N^* \setminus H^*(\mathcal{R}^*)$.*

Proof. The proof proceeds by steps. **Step 1.** Consider any maximal set $M \subset N^* \setminus H^*(\mathcal{R}^*)$ of nodes connected through non-fundamental links and let $M^+ \subset H^*(\mathcal{R}^*)$ be the set of nodes that have non-fundamental links to nodes in M. By Lemma 1, there is a $d > 0$ so

that $M, M^+ \subset N_d$. Denote by M^{++} the set of nodes in $N_d \cap H^*(\mathcal{R}^*)$ with fundamental links into M. Since the nodes in M are connected through non-fundamental links, there is a fundamental link from any node $i \in M^{++}$ to any node in M. Thus, any node in M^{++} must also be adjacent to any node in M^+, so $M^+ \cup M^{++}$ is a clique. **Step 2.** Consider the DAG $\bar{\mathcal{G}} = (N, \bar{G})$, where $N = M \cup M^+ \cup M^{++}$ and $\bar{G}$ is identical to R^* restricted on N. By construction, this DAG is perfect. Hence, Corollary 1 from Spiegler (2019) implies that there exists a DAG $\bar{\mathcal{G}}^+$ in which the clique $M^+ \cup M^{++}$ is ancestral and that is equivalent to $\bar{\mathcal{G}}$. We choose such a $\bar{\mathcal{G}}^+$ with the property that the ordering of the nodes in $M^+ \cup M^{++}$ is the same as in $\bar{\mathcal{G}}$ (this is possible since $M^+ \cup M^{++}$ is a clique, and all links between nodes $M^+ \cup M^{++}$ and nodes in M point towards the latter one). Consider now the DAG $\mathcal{G}$ that is identical to $\mathcal{R}^*$ except that $\bar{\mathcal{G}}$ is replaced by $\bar{\mathcal{G}}^+$. We show that there are no cycles or v-colliders in $\mathcal{G}$ so that it is equivalent to $\mathcal{R}^*$. Consider any node $i \in N_{d-1} \cup N_d$ that is outside $M \cup M^+ \cup M^{++}$ and that has a fundamental link into a node in M. Since the nodes in M are connected through non-fundamental links, node i has a fundamental link into every node in M (otherwise, i would belong to M, a contradiction). This rules out v-colliders. Any link between a node in N_d and a node in N_{d+1} points into the latter one. Hence, by construction, there cannot be cycles or v-colliders in $\mathcal{G}$. We obtain $\mathcal{G}^*$ by performing the same changes for any maximal set $M \subset N^* \setminus H^*(\mathcal{R}^*)$ of nodes connected by non-fundamental links in $\mathcal{R}^*$. □

Proof of Proposition 1.7. First, we show the "if"-statement. Assume that the agent's subjective model $\mathcal{R}$ contains all the nodes in $H^*(\mathcal{R}^*)$. Consider the DAG $\mathcal{G}^* \in \mathcal{E}$ in which all links with one end in $H^*(\mathcal{R}^*)$ and the other in $N^* \setminus H^*(\mathcal{R}^*)$ point from $H^*(\mathcal{R}^*)$ to $N^* \setminus H^*(\mathcal{R}^*)$. By Lemma 6, this DAG exists. From Proposition 1.6 it follows that $p_{\mathcal{G}^*}(x_{H^*(\mathcal{R}^*)}) = p(x_{H^*(\mathcal{R}^*)})$ for all distributions $p(x_{N^*}) \in \Delta(X_{N^*})$. Consider the subgraph $\mathcal{G} = (G, N)$ where G equals G^* restricted on N. Since none of the nodes in $N \setminus H^*(\mathcal{R}^*)$ impacts on any node in $H^*(\mathcal{R}^*)$, we have $p_{\mathcal{G}}(x_{H^*(\mathcal{R}^*)}) = p_{\mathcal{G}^*}(x_{H^*(\mathcal{R}^*)})$ for all $p(x) \in \Delta(X)$. By construction, the DAGs $\mathcal{R}$ and $\mathcal{G}$ are equivalent so that we have $p_{\mathcal{R}}(x_{H^*(\mathcal{R}^*)}) = p_{\mathcal{G}}(x_{H^*(\mathcal{R}^*)}) = p_{\mathcal{G}^*}(x_{H^*(\mathcal{R}^*)}) = p(x_{H^*(\mathcal{R}^*)})$ for all distributions $p(x_{N^*}) \in \Delta(X_{N^*})$, which proves the "if"-statement. Next, we show the "only if"-statement. Assume that there is one node $i \in H^*(\mathcal{R}^*)$ that is not in the agent's subjective model. This node is on a fundamental active path τ between the action node 0 and the output node n. We then can find a probability distribution $p(x_{N^*}) \in \Delta(X_N^*)$ so that $p_{\mathcal{R}}(x_n \mid x_0) \neq p(x_n \mid x_0)$. Let k be the k'th node in τ. Consider a probability distribution with the following properties: $p(x_j \mid x_{R^*(j)}) = p(x_j)$ for all nodes $j \notin \tau$ that are between the nodes 0 and n, and $p(x_k \mid x_{R^*(k)}) = p(x_k \mid x_{k-1})$. Clearly, such a distribution can have the desired property. □

Proof of Corollary 3. Denote $H^*(\mathcal{R}_1) = H^*(\mathcal{R}_2) = H$. By Proposition 1.7, there exists a DAG $\mathcal{R}_1^{[1]}$ that is equivalent to $\mathcal{R}_1$ and in which all links between any node $i \in H$ and any node $j \in N_1 \setminus H$ is turned away from i. Thus, we have

$$p_{\mathcal{R}_1}(x_H) = \sum_{x_{N_1 \setminus H} \in X_{N_1 \setminus H}} p_{\mathcal{R}_1}(x_{N_1}) = \sum_{x_{N_1 \setminus H} \in X_{N_1 \setminus H}} p_{\mathcal{R}_1^{[1]}}(x_{N_1}) = p_{\mathcal{R}_1^{[1]}}(x_H). \tag{A.19}$$

Note that for all $i \in H$ we have that $R_1^{[1]}(i) \subset H$. Consider the restriction of $R_1^{[1]}$ on H, $R_1^{[H]}$. We then have

$$p_{\mathcal{R}_1^{[1]}}(x_H) = \prod_{i \in H} p(x_i \mid x_{R_1^{[1]}(i)}) = \prod_{i \in H} p(x_i \mid x_{R_1^{[H]}(i)}) = p_{\mathcal{R}_1^{[H]}}(x_H). \tag{A.20}$$

Define $\mathcal{R}_2^{[1]}$ and $\mathcal{R}_2^{[H]}$ just like $\mathcal{R}_1^{[1]}$ and $\mathcal{R}_1^{[H]}$. By assumption, the link $iR_1^{[H]}j$ is in $R_1^{[H]}$ if and only if we have $iR_2^{[H]}j$ or $jR_2^{[H]}i$. Thus, $\mathcal{R}_1^{[H]}$ and $\mathcal{R}_2^{[H]}$ have the same skeleton. Since $\mathcal{R}_1$ and $\mathcal{R}_2$ are perfect, so are $\mathcal{R}_1^{[H]}$ and $\mathcal{R}_2^{[H]}$. Hence $\mathcal{R}_1^{[H]}$ and $\mathcal{R}_2^{[H]}$ are equivalent, so that

$$p_{\mathcal{R}_1^{[H]}}(x_H) = p_{\mathcal{R}_2^{[H]}}(x_H). \tag{A.21}$$

From the equations (A.19) to (A.21), we get $p_{\mathcal{R}_1}(x_H) = p_{\mathcal{R}_2}(x_H)$, which implies the result. □

1.8.6 Risk and Incentives

To study the relationship between risk and incentives, the literature typically uses a setting with continuous actions, normally distributed output, and exponential utility so that the optimal contract is linear. To properly apply our framework, we consider a setting with discrete actions and outputs that captures the negative relationship between risk and incentives.

Let there be a binary action $a \in \{0, 1\}$ and three equidistant output levels, y_L, y_M, y_H with $y_H > y_M > y_L > 0$. The level of risk is indexed by a parameter $\xi \in [0, \bar{\xi}]$. The production function is $p(y_L \mid a) = \beta_L(\xi) - \beta a$, $p(y_M \mid a) = \beta_M(\xi)$, and $p(y_H \mid a) = \beta_H(\xi) + \beta a$, where $\beta_L(\xi) = \beta_H(\xi)$ for all ξ. An increase in risk ξ shifts probability mass from the medium output y_M to the extreme outputs y_L and y_H, i.e., $\beta_L'(\xi) = \beta_H'(\xi) = \varepsilon$ for some $\varepsilon > 0$ and $\beta_M'(\xi) = -2\varepsilon$. The agent has a piecewise linear utility function $u(w) = w$ for $w \geq 0$, and $u(w) = \lambda w$ with $\lambda > 1$ for $w < 0$. Her reservation utility is $\bar{U} = 0$.

We now fit the marketer example from Subsection 1.3.2 to the present setting. The objective causal model is given by $\mathcal{R}^*$ on the left of Figure 1.2, while the agent's subjective model is given by $\mathcal{R}$ on the right of this figure. We use our usual parametrization, except for the output. The probability of low, middle, and high output conditional on x_1 and x_2 is given by

$$p(y_H \mid x_1, x_2) = \beta_3^H(\xi) + \beta_{13}(\xi)x_1 + \beta_{23}(\xi)x_2, \tag{A.22}$$

$$p(y_M \mid x_1, x_2) = \beta_3^M(\xi), \tag{A.23}$$

$$p(y_L \mid x_1, x_2) = \beta_3^L(\xi) - \beta_{13}(\xi)x_1 - \beta_{23}(\xi)x_2. \tag{A.24}$$

The level of risk ξ changes the importance of consumer information and reputation for the final output. The larger the risk, the more important are these two factors to obtain a high rather than a small output. We capture this by assuming

$$\beta_{13}(\xi) = \bar{\beta}_{13}\left(1 + \frac{\xi}{\beta_{01}\bar{\beta}_{13}}\right) \quad \text{and} \quad \beta_{23}(\xi) = \bar{\beta}_{23}\left(1 + \frac{\xi}{\mid \beta_{02} \mid \bar{\beta}_{23}}\right) \tag{A.25}$$

for two values $\bar{\beta}_{13}, \bar{\beta}_{23} > 0$ with $\beta_{01}\bar{\beta}_{13} + \beta_{02}\bar{\beta}_{23} = \beta$. We choose the functions $\beta_3^H(\xi)$, $\beta_3^M(\xi)$ and $\beta_3^L(\xi)$ so that the objective probability model generates the production function from above.[15]

Proposition 1.9 (Risk and Incentives)**.** *Consider the marketer example of this subsection.*

(a) Suppose the agent's subjective model equals $\mathcal{R}^*$. The expected wage payment needed to implement $\alpha = 1$ then increases in risk ξ, and there exists an interval $[c_L, c_H]$ so that if $c \in (c_L, c_H)$, then for some $\xi^* \in (0, \bar{\xi})$ the optimal equilibrium contract implements $\alpha = 1$ if $\xi < \xi^*$ and $\alpha = 0$ if $\xi > \xi^*$.

(b) Suppose the agent's subjective model equals $\mathcal{R}$. The expected wage payment needed to implement $\alpha = 1$ then decreases in risk ξ if the slope $\beta_L'(\xi) = \beta_H'(\xi) = \varepsilon$ is small enough. In this case, there is an interval $[c_L, c_H]$ so that if $c \in (c_L, c_H)$, then for some $\xi^* \in (0, \bar{\xi})$ the optimal equilibrium contract implements $\alpha = 0$ if $\xi < \xi^*$ and $\alpha = 1$ if $\xi > \xi^*$.

[15] Specifically, we derive $\beta_3^H(\xi)$ and $\beta_3^L(\xi)$ from $\beta_H(\xi) = \beta_3^H(\xi) + \beta_1\beta_{13}(\xi) + \beta_2\beta_{23}(\xi)$ and $\beta_L(\xi) = \beta_3^L(\xi) - \beta_1\beta_{13}(\xi) - \beta_2\beta_{23}(\xi)$. Since $\beta_H(\xi) = \beta_L(\xi)$ for all ξ, we have $\beta_3^M(\xi) = 1 - 2[\beta_3^H(\xi) + \beta_1\beta_{13}(\xi) + \beta_2\beta_{23}(\xi)]$.

Below we provide the proof of Proposition 1.9. We explain why part (a) holds. When the agent has rational expectations, the IC that ensures high effort equals

$$\beta(u(w_H) - u(w_L)) \geq c, \tag{A.26}$$

and the optimal wage schedule that implements high effort is given by

$$w(y_L) = -\frac{1}{2\lambda\beta}c, \; w(y_M) = 0, \text{ and } w(y_H) = \frac{1}{2\beta}c. \tag{A.27}$$

Note that a change in risk ξ affects neither the optimal wage schedule, nor the incentive compatibility constraint in (A.26). In terms of effort incentives, the effect of risk on the importance of consumer information and reputation cancel each other out. However, an increase in risk exposes the agent to more variation, so that she requires a higher risk-premium. Hence, when the principal implements high effort, his expected payment to the agent under the optimal contract increases in risk. Therefore, there exists an interval of cost levels $[c_L, c_H]$, so that if $c \in (c_L, c_H)$, the optimal equilibrium contract implements high effort if and only if the level of risk is sufficiently small. We thus obtain a negative relationship between risk and incentives.

Next, consider part (b). If the agent does not take reputation into account, an increase in risk appears to her as an increase in the productivity of her effort, as the association between consumer information and sales becomes stronger. The IC that ensures high effort now equals

$$\beta_{01}\beta_{13}(\xi)(u(w_H) - u(w_L)) \geq c. \tag{A.28}$$

Recall that $\beta_{13}(\xi)$ increases in ξ. Hence, an increase in risk ξ relaxes this IC. The optimal wage schedule that implements $\alpha = 1$ is now given by

$$w(y_L) = -\frac{\beta_H(\xi) + \beta - \beta_{01}\beta_{13}(\xi)}{\lambda(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)}c, \; w(y_M) = 0, \text{ and } w(y_H) = \frac{\beta_L(\xi) - \beta + \beta_{01}\beta_{13}(\xi)}{(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)}c. \tag{A.29}$$

A change in risk now has two countervailing effects on the expected payment when the principal implements high effort. It again increases the risk premium that the agent requires, but it also relaxes the incentive compatibility constraint. Which effect dominates depends on the probability model and the utility function. If the slope $\beta'_L(\xi) = \beta'_H(\xi) = \varepsilon$ is small enough, an increase in risk reduces the expected payment to the agent at all risk levels $\xi \in [0, \bar{\xi}]$. We then obtain a positive relationship between risk and incentives: For an interval of cost levels $[c_L, c_H]$, if $c \in (c_L, c_H)$, the optimal equilibrium contract implements high effort if the level of risk is sufficiently large, and otherwise low effort through a fixed

wage.

Proof of Proposition 1.9. We first prove statement (a). For this, we derive the optimal contract under the objective model $\mathcal{R}^*$ that implements high effort. For convenience, we abbreviate $w_H = w(y_H)$, $w_M = w(y_M)$, and $w_L = w(y_L)$. Standard arguments show that both IC and PC must be binding at the optimal contract, and that $w_L < 0$ and $w_H > 0$ at the optimum. Assume for the moment that $w_M \geq 0$ under the optimal contract. The IC is then

$$\beta(w_H - \lambda w_L) = c, \tag{A.30}$$

and the PC equals

$$(\beta_H(\xi) + \beta)w_H + \beta_M(\xi)w_M + (\beta_L(\xi) - \beta)\lambda w_L = 0. \tag{A.31}$$

From the IC we get

$$w_H = \frac{c}{\beta} + \lambda w_L, \tag{A.32}$$

We plug this into the PC, solve for w_M, and get

$$w_M = -\frac{\beta_H(\xi)}{\beta_M(\xi)\beta}c - \frac{\beta_L(\xi) + \beta_H(\xi)}{\beta_M(\xi)}\lambda w_L. \tag{A.33}$$

The expected wage payment of the principal when he implements $\alpha = 1$ equals

$$\mathbb{E}[w \mid \alpha = 1] = (\beta_H(\xi) + \beta)w_H + \beta_M(\xi)w_M + (\beta_L(\xi) - \beta)w_L. \tag{A.34}$$

Using the results from above, we can write the expected wage payment as

$$\mathbb{E}[w \mid \alpha = 1] = c - (\beta_L(\xi) - \beta)(\lambda - 1)w_L. \tag{A.35}$$

The optimal wage w_L minimizes this term subject to the constraint that w_M in (A.33) remains weakly positive. The solution implies that $w_M = 0$, and $w(y_L) = -\frac{1}{2\lambda\beta}c$ as well as $w(y_H) = \frac{1}{2\beta}c$. We obtain the same result when we go through the same steps while assuming $w_M \leq 0$. With this we can compose the expected wage payment $\mathbb{E}[w \mid \alpha = 1]$ and obtain

$$\frac{\partial \mathbb{E}[w \mid \alpha = 1]}{\partial \xi} = \frac{\varepsilon}{2\beta}c - \frac{\varepsilon}{2\lambda\beta}c > 0. \tag{A.36}$$

Hence, the expected wage payment to implement $\alpha = 1$ strictly increases in risk. The expected wage payment to implement $\alpha = 0$ is zero for all risk levels. This yields us statement (a).

Next, we prove statement (b). We first derive the agent's beliefs about the production function at $\alpha = 1$. As in the proof of Proposition 1.2, we calculate $p(x_2 = 1 \mid x_1 = 1)$ and $p(x_2 = 1 \mid x_1 = 0)$. At $\alpha = 1$, we have $p(x_2 = 1 \mid x_1 = 1) = p(x_2 = 1 \mid x_1 = 0) = \beta_2 + \beta_{02}$, and thus

$$\begin{aligned}
p(y_H \mid x_1 = 1) &= \beta_3^H(\xi) + \beta_{13}(\xi) + (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.37)}\\
p(y_H \mid x_1 = 0) &= \beta_3^H(\xi) + (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.38)}\\
p(y_M \mid x_1 = 1) &= p(y_M \mid x_1 = 0) = \beta_3^M(\xi), && \text{(A.39)}\\
p(y_L \mid x_1 = 1) &= \beta_3^L(\xi) - \beta_{13}(\xi) - (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.40)}\\
p(y_L \mid x_1 = 0) &= \beta_3^L(\xi) - (\beta_2 + \beta_{02})\beta_{23}(\xi). && \text{(A.41)}
\end{aligned}$$

From this, we can derive the agent's beliefs about the production function at $\alpha = 1$ as

$$\begin{aligned}
p_{\mathcal{R}}(y_H \mid a = 1; \alpha = 1) &= \beta_3^H(\xi) + (\beta_1 + \beta_{01})\beta_{13}(\xi) + (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.42)}\\
p_{\mathcal{R}}(y_H \mid a = 0; \alpha = 1) &= \beta_3^H(\xi) + \beta_1\beta_{13}(\xi) + (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.43)}\\
p_{\mathcal{R}}(y_M \mid a = 1; \alpha = 1) &= p_{\mathcal{R}}(y_M \mid a = 0; \alpha = 1) = \beta_3^M(\xi), && \text{(A.44)}\\
p_{\mathcal{R}}(y_L \mid a = 1; \alpha = 1) &= \beta_3^L(\xi) - (\beta_1 + \beta_{01})\beta_{13}(\xi) - (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.45)}\\
p_{\mathcal{R}}(y_L \mid a = 0; \alpha = 1) &= \beta_3^L(\xi) - \beta_1\beta_{13}(\xi) - (\beta_2 + \beta_{02})\beta_{23}(\xi). && \text{(A.46)}
\end{aligned}$$

At $\alpha = 1$, the IC is therefore given by

$$\beta_{01}\beta_{13}(\xi)(u(w_H) - u(w_L)) \geq c. \tag{A.47}$$

The rest of the proof proceeds as in the proof of statement (a). We derive the equilibrium contract that implements $\alpha = 1$ at lowest cost to the principal when the agent's subjective model is given by $\mathcal{R}$. Assume that we have $w_M \geq 0$ at this contract. From the IC, we get

$$w_H = \frac{c}{\beta_{01}\beta_{13}(\xi)} + \lambda w_L, \tag{A.48}$$

and from the PC we get that

$$w_M = -\frac{\beta_H(\xi) + \beta - \beta_{01}\beta_{13}(\xi)}{\beta_M(\xi)\beta_{01}\beta_{13}(\xi)} - \frac{\beta_L(\xi) + \beta_H(\xi)}{\beta_M(\xi)}\lambda w_L. \tag{A.49}$$

With this, we can calculate the expected wage payment under the optimal equilibrium contract that implements $\alpha = 1$ as

$$\mathbb{E}[w \mid a = 1; \mathcal{R}] = c - (\beta_L(\xi) - \beta)(\lambda - 1)w_L. \tag{A.50}$$

The optimal wage w_L minimizes this term subject to the constraint that w_M in (A.49) remains weakly positive. The solution implies that $w_M = 0$ as well as

$$w_L = -\frac{\beta_H(\xi) + \beta - \beta_{01}\beta_{13}(\xi)}{\lambda(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)}c \text{ and } w_H = \frac{\beta_L(\xi) - \beta + \beta_{01}\beta_{13}(\xi)}{(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)}c. \tag{A.51}$$

We obtain the same result when we go through the same steps while assuming $w_M \leq 0$. We then can compose the expected wage payment at the optimal equilibrium contract that implements $\alpha = 1$ as

$$\mathbb{E}[w \mid a = 1; \mathcal{R}] = \frac{(\lambda - 1)(\beta_H(\xi) + \beta)(\beta_L(\xi) - \beta) + (\lambda + 1)\beta_{01}\beta_{13}(\xi)}{\lambda(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)}. \tag{A.52}$$

We differentiate this expression with respect to risk ξ and find

$$\lim_{\varepsilon \to 0} \frac{\partial \mathbb{E}[w \mid a = 1; \mathcal{R}]}{\partial \xi} = -\frac{\lambda(\lambda - 1)(\beta_H(\xi) + \beta_L(\xi))(\beta_H(\xi) + \beta)(\beta_L(\xi) - \beta)}{[\lambda(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)]^2} < 0. \tag{A.53}$$

Hence, if ε is sufficiently small, the expected wage payment needed to implement $\alpha = 1$ decreases in risk ξ. The rest of the proof of statement (b) proceeds in the same way as for statement (a). □

Chapter 2

Strategic Interpretations

2.1 Introduction

In the simplest textbook model of strategic communication, originated by Crawford and Sobel (1982), a "sender" privately observes a state of Nature and chooses a costless message from some given message space. Then, a "receiver" observes the message and takes an action that affects both parties' payoffs. A hallmark of this conventional approach is that messages have no intrinsic meaning; their content - namely, their statistical relation with the underlying state - is established in Nash equilibrium of the sender-receiver game. According to the standard steady-state interpretation of this solution concept, the receiver has access to a "dataset" that fully reveals the statistical relation between states and messages.

In this paper we revisit the basic sender-receiver model and relax the assumption that the receiver is fully capable of interpreting equilibrium messages. We focus on settings in which the receiver has two available actions, y and n. In each state of Nature, exactly one of these actions is appropriate. The prior probability of the states for which y is the appropriate action is $\pi < \frac{1}{2}$. The receiver's sole objective is to select the appropriate action. This familiar setting is borrowed from Glazer and Rubinstein, Glazer and Rubinstein (2004, 2006) or Kamenica and Gentzkow (2011). For most of the paper, we follow these papers by also assuming that the sender always wants the receiver to play y (but we also examine an alternative, "zero-sum" specification).

By default, our receiver lacks access to any data regarding the state-message mapping, and therefore cannot decipher messages by himself. He is like a tourist in a foreign country who

does not understand its language or cultural codes. However, if an "interpreter" handed him a "*dictionary*" containing data regarding the statistical mapping from states to the sender's messages, he would have some ability to interpret the message he receives.

Our model makes room for the *strategic* supply of such dictionaries. The sender himself - or a *third party* who acts as an *interpreter* on the sender's behalf - chooses a dictionary from some feasible set. He can condition the dictionary on the state and the message. Thus, different messages may be accompanied by different dictionaries, and the same message may be paired with different dictionaries in different states. Each dictionary provides *credible*, yet possibly selective statistical data regarding the sender's state-message mapping (given by the sender's strategy). The receiver uses this data to update his belief given the message. Crucially, our basic model assumes that the receiver lacks any other means for extracting the meaning of messages (we relax this assumption in Section 4). Consequently, he *does not draw any inferences from the provided dictionary itself*, since this would require some data regarding the *joint* distribution of messages, dictionaries and states - data the receiver does not have.[1] Consequently, the sender can manipulate the receiver's beliefs beyond what is feasible under rational expectations.

Strategic interpretation of messages - in the sense of providing selective statistical data about their meaning - is pervasive in real-life situations, whether the messages are cheap talk or hard-information disclosures. Consider an employee who wants to exert effort only when sufficiently sure he is not about to be fired. He is summoned to the General Manager's office to hear about his prospects at the company. After the meeting is over, the HR manager (who was present at the meeting) explains that when the GM says to an employee "you have a future in the company", this means a 50% chance of keeping his job. This is an interpretation of the GM's verbal message. It is selective because it ignores other aspects of the GM's communication, e.g. his body language. Alternatively, the HR's interpretation could focus on the latter: "The GM's handshake was feeble; this is definitely bad news".

Another example involves a tenure case that is brought in front of a university promotions committee. Although the candidate submits his CV, committee members outside his discipline cannot decipher the connection between the candidate's quality and indicators such as the number of publications, conference lectures or supervised students. The candidate's department chair will offer an interpretation by providing statistical data about researchers in comparable departments (including their subsequent academic performance,

[1] A similar form of bounded rationality is documented in Jin et al. (2019), who find that in a laboratory game of voluntary disclosure, receivers do not make correct inferences from no disclosure.

which indicates their "true quality"). If the chair's objective is misaligned with the university committee's, the data he provides may be strategically selective. In a similar vein, imagine a foreign candidate for a graduate program. The candidate submits his grade transcript, yet the admission committee does not know the grades' meaning. A faculty member writing a recommendation letter on the candidate's behalf may provide such an interpretation, by describing the grade distribution for a selected subset of courses.

Finally, suppose the sender is a political party and the receiver is a representative voter. The party's message is multi-dimensional, where each component describes public pronouncements by a different party member. A political commentator interprets the party's message in some media outlet. He does so by providing historical data about the match between the public pronouncements of selected party organs and the underlying reality.

These are all examples of selective interpretations where the receiver is presented with partial statistics about the sender's state-dependent, multi-dimensional message. These interpretations can be strategic when the interpreter's interests are misaligned with the receiver's. We analyze the sender's choice of messages when he takes their subsequent strategic interpretation into account. For instance, the way a political party structures the public statements made by its members will be shaped by its expectation of how a media outlet that is biased in its favor will interpret these statements.

One could argue that in these examples, the statistical data the interpreter provides need not be perfectly credible or unbiased. However, because they are quantitative and verifiable, they are more likely to be credible than cheap-talk messages like "you have a future in the company". At any rate, we abstract from this consideration; our analytical task is to quantify the effect of strategic provision of *cheap-talk* messages and their interpretation on the sender's ability to attain his objective, assuming perfect credibility of the statistical data these interpretations involve. In the course of this paper, we will consider various kinds of partial statistics that strategic interpretations can entail.

Preview of the analysis

We present our basic model in Section 2, where we define a dictionary as a non-empty subset of the components of a K-dimensional message. The dictionary enables the receiver to learn the state-dependent joint distribution of these components. We assume that the interpreter's preferences fully coincide with the sender's. For expositional convenience, our formal exposition regards them as a *single* player who *commits* to a state-dependent joint distribution over messages and dictionaries. Neither of these two assumptions is necessary

for our main findings. (In our informal description, we occasionally refer to the interpreter as a distinct agent who shares the sender's preferences.)

In Section 3, we present our main result, which characterizes the maximal probability of persuasion as a function of π and K. In particular, we show that the sender can attain full persuasion, as long as π is above a cutoff $\pi^*(K)$ given by a simple formula that makes use of Sperner's Theorem and decays quickly with K.

Our assumption that the receiver draws no inferences from the dictionary he is given raises natural questions. First, does the dictionary itself convey information about the underlying state? The answer is negative: The sender-optimal strategy we construct has the property that the distribution over dictionaries is state-independent. Second, would the receiver be "suspicious" of a dictionary that does not cover all message components? We address this question in Section 4, while insisting on sender strategies that induce a state-independent dictionary distribution.

In Section 4.1, we perturb the model by assuming that the sender has a lexicographically secondary preference for small dictionaries. We also introduce a refinement of the sender's strategy: if the sender's interests were aligned with the receiver's, he would want to play a strategy that induces the same observed distribution over dictionaries. Thus, if the receiver had independent access to data about the distribution of dictionaries, he could reconcile the observed use of selective dictionaries with a benevolent sender. Under this refinement, we show that full persuasion is attainable if and only if $\pi \geq 1/(K+1)$. The sender's strategy only interprets *single* message components.

In Section 4.2, we modify the definition of dictionaries. When a dictionary $D \subseteq \{1, ..., K\}$ is provided, this now means that the receiver learns the state-dependent distribution of m_D *as well as* the state-dependent distribution of $m_{\{1,...,K\}\setminus D}$. Thus, the interpreter is forced to provide statistical data about the behavior of *all* message components, though in a format that can break them into two disjoint sets. Under a mild assumption on how the receiver extrapolates a belief from these pieces of data, we show that full persuasion is attainable whenever π exceeds a cutoff that decays quickly with K. The lesson from these two variants of our basic model is that strategic interpretation can produce effective persuasion without generating excessive "suspicion" regarding its selectivity.

Section 5 picks up the theme of Section 4.2 and present an example that illustrates a richer notion of dictionaries, which involves data about other slices of the joint state-message distribution. We show how this richer specification can enhance the sender's ability to

attain full persuasion. In Section 6 we perform partial analysis of our basic model when the two parties have diametrically opposed preferences. We discuss related literature in Section 7.

2.2 A Model

There are two players, a sender and a receiver. The sender observes a state of Nature $\theta \in \Theta = \{Y, N\}$. The receiver does not observe the state but needs to take an action a, which can be either "yes" (denoted y) or "no" (denoted n). Players' payoffs take values in $\{0, 1\}$. The receiver's payoff is 1 if either $< a = y$ and $\theta = Y >$ or $< a = n$ and $\theta = N >$, and 0 otherwise. In contrast, the sender's payoff is 1 if and only if $a = y$, and 0 otherwise.

The players' common prior belief over Θ assigns probability $\pi < \frac{1}{2}$ to state Y. Hence, the receiver's ex-ante optimal action is n. However, the sender can influence the receiver's belief and persuade him to play y. He commits to a strategy that maps each state to a distribution over *reports,* where a report is a pair (m, D) such that:

(i) $m = (m_1, ..., m_K) \in M^K$ is a K-dimensional *message*, where $K \geq 1$ and $|M| \geq 2$. In all the examples we use in the paper, $M = \{0, 1\}$.

(ii) $D \in 2^{\{1,...,K\}} \backslash \{\emptyset\}$ is a *dictionary.*

Thus, the sender's strategy is a function $\sigma : \Theta \to \Delta\left(M^K \times 2^{\{1,...,K\}} \backslash \{\emptyset\}\right)$. The commitment assumption is made for expositional simplicity; as we shall see, our results regarding full persuasion are insensitive to it. The assumption that $|\Theta| = 2$ could be replaced with the weaker assumption that there is a function $f : \Theta \to \{n, y\}$ such that the receiver's payoff is 1 if and only if $a = f(\theta)$, and 0 otherwise. The probability with which the sender plays the report (m, D) in state θ is denoted $\sigma(m, D \mid \theta)$. With slight abuse of notation, define $\sigma(m \mid \theta) = \sum_D \sigma(m, D \mid \theta)$ and $\sigma(D \mid \theta) = \sum_m \sigma(m, D \mid \theta)$. We refer to $(\sigma(m \mid \theta))$ as the *message strategy* and to $(\sigma(D \mid m, \theta))$ as the *interpretation strategy.*

The role of dictionaries is to grant the receiver "partial access" to the statistical regularities of the sender's strategy. When the receiver observes the report (m, D), he learns the conditional probabilities $(\sigma(m_D \mid \theta))_{\theta \in \Theta}$, where $m_D = (m_k)_{k \in D}$ and

$$\sigma(m_D \mid \theta) = \sum_{m' \mid m'_D = m_D} \sigma(m' \mid \theta)$$

That is, the receiver learns how the message components in D - *and nothing but them* - are distributed conditional on the state. He *cannot* draw any statistical inferences from the message components $m_{\{1,...,K\}\setminus D}$ or the dictionary D itself. We will revisit this assumption in the sequel. Note that in any report (m, D), D must be a *non-empty* subset of $\{1, ..., K\}$; that is, the sender is obliged to provide *some* interpretation of the message.

Upon receiving a report (m, D), the receiver updates his belief according to the following expression:

$$\widetilde{\Pr}(\theta = Y \mid m, D) = \frac{\pi \cdot \sigma(m_D \mid \theta = Y)}{\pi \cdot \sigma(m_D \mid \theta = Y) + (1 - \pi) \cdot \sigma(m_D \mid \theta = N)} \tag{2.1}$$

Compare this with the correct, rational-expectations posterior probability of Y conditional on (m, D):

$$\Pr(\theta = Y \mid m, D) = \frac{\pi \cdot \sigma(m, D \mid \theta = Y)}{\pi \cdot \sigma(m, D \mid \theta = Y) + (1 - \pi) \cdot \sigma(m, D \mid \theta = N)} \tag{2.2}$$

The receiver best-replies to the subjective posterior belief given by (2.1), breaking ties in favor of the sender. Equivalently, faced with a report (m, D), he computes its subjective likelihood ratio

$$\rho_\sigma(m, D) = \frac{\sum_{m' | m'_D = m_D} \sigma(m' \mid \theta = Y)}{\sum_{m' | m'_D = m_D} \sigma(m' \mid \theta = N)} \tag{2.3}$$

and chooses $a = y$ if and only if $\rho_\sigma(m, D) \geq (1 - \pi)/\pi$.

The sender chooses his strategy under the assumption that the receiver best-replies to the belief given by (2.1). Our main question is: What is the maximal probability of $a = y$ that the sender can attain?

Our model of how the receiver forms beliefs is motivated by the steady-state view of equilibrium behavior, whereby the sender's strategy σ describes a long-run statistical relation between states and reports. The receiver moves once, against the background of a large dataset consisting of many realizations of $(\theta, m_1, ..., m_K, D)$ resulting from previous interactions between the sender with different identical receivers. The dataset can be visualized as a large spreadsheet, where each column represents one of the variables $\theta, m_1, ..., m_K, D$, and each row represents an observation (an independent draw from the joint distribution over states and reports). Rational expectations correspond to having full access to this dataset. Our model relaxes this assumption and assumes that the receiver is granted access to a subset of columns represented by D. The receiver can only rely on the accessed data for drawing inferences.

Example 1

To illustrate our notion of dictionaries and how the receiver reacts to them, suppose that $K = 4$. Assume $\sigma(m \mid Y)$ is uniform over $(1,1,1,1)$ and $(0,0,0,0)$, while $\sigma(m \mid N)$ is uniform over $(1,1,1,1)$, $(1,0,1,0)$ and $(1,0,0,1)$.

Suppose the sender accompanies the message $(1,0,1,0)$ with the dictionary $D = \{1,3\}$. This dictionary provides the receiver with data about the state-dependent distribution of (m_1, m_3). In particular, he learns that the pattern $(1,*,1,*)$ occurs with probability $\frac{1}{2}$ in state Y and with probability $\frac{2}{3}$ in state N.[2] Therefore,

$$\widetilde{\Pr}(\theta = Y \mid (1,0,1,0), \{1,3\}) = \frac{\pi \cdot \frac{1}{2}}{\pi \cdot \frac{1}{2} + (1-\pi) \cdot \frac{2}{3}} = \frac{3\pi}{4-\pi}$$

By comparison, the rational-expectations posterior on Y given $m = (1,0,1,0)$ is zero (independently of the dictionary that accompanies this message).

Note that the message $(1,1,1,1)$ is sent with positive probability in *both* states. Suppose that in state Y the sender accompanies this message with the dictionary $D = \{1,2,3\}$. The receiver then learns that the pattern $(1,1,1,*)$ occurs with probability $\frac{1}{2}$ in state Y and with probability $\frac{1}{3}$ in state N. Hence,

$$\widetilde{\Pr}(\theta = Y \mid (1,1,1,1), \{1,2,3\}) = \frac{\pi \cdot \frac{1}{2}}{\pi \cdot \frac{1}{2} + (1-\pi) \cdot \frac{1}{3}} = \frac{3\pi}{2+\pi}$$

Suppose next that in state N the sender accompanies the message $(1,1,1,1)$ with the dictionary $D = \{3\}$. Then

$$\widetilde{\Pr}(\theta = Y \mid (1,1,1,1), \{3\}) = \frac{\pi \cdot \frac{1}{2}}{\pi \cdot \frac{1}{2} + (1-\pi) \cdot \frac{2}{3}} = \frac{3\pi}{4-\pi}$$

Thus, by varying the dictionary across states, the same message induces the receiver to hold a different belief in each state. In contrast, if the receiver had rational expectations, then independently of the dictionary, his posterior on Y given $m = (1,1,1,1)$ would be $\frac{3\pi}{2+\pi}$ in *both* states. □

We close this section with comments on a few aspects of our model.

[2]The notation $(1,*,1,*)$ stands for all messages m for which $m_1 = m_3 = 1$.

Multi-dimensional messages

The multi-dimensionality of messages has a few interpretations. First, different components of m may represent different modes of communication (verbal statements, voice intonation). When the sender is an organization, different components represent utterances by different organs (party whip, corporate executive, spokesperson). Finally, the state itself can be multi-dimensional (this requires $|\Theta| > 2$), such that each message component corresponds to a different state dimension.

Rational expectations and the full dictionary

Note that the full dictionary $D = \{1, ..., K\}$ does *not* automatically endow the receiver with rational expectations. The reason is that rational expectations mean that the receiver knows the sender's entire reporting strategy, whereas the full dictionary only enables him to learn the message strategy. However, if the interpretation strategy happens to be measurable with respect to messages (i.e. $\sigma(D \mid m) \equiv \sigma(D \mid m, \theta)$), accompanying a message with the full dictionary will enable the receiver to update his belief as if he had rational expectations.

The "redacted message" metaphor

Our model could be alternatively described as follows. When the sender sends a message, he selectively "redacts" parts of that message, such that the receiver gets to observe only the unredacted parts. The belief-formation rule (2.1) means that the receiver takes into account the sender's pre-redaction message strategy but ignores the redaction strategy (and therefore draws no inference from the redacted components).

We find this "selective redaction" description less appealing because it lacks a concrete story for how the receiver forms correct expectations about the sender's message strategy but not about the redaction strategy. In contrast, our original description of D as a representation of selective statistical data regarding the sender's strategy entails an explicit mechanism for this dichotomy: The receiver can only base his beliefs on the statistical data provided to him by the sender.

More importantly, our description opens the door for *other types of dictionaries* that correspond to other kinds of statistical data that the sender can transmit to the receiver. We illustrate this idea in Sections 4.2 and 5, where we allow the sender to provide multiple "datasets" that record different slices of the joint state-message distribution, and show how this richer notion of dictionaries affects the sender's problem. These extensions of our basic

model go beyond the scope of the "redaction" metaphor.

Who interprets the messages?

Given that we model the situation as a two-player game, a literal interpretation of our model would be that the sender interprets his own messages. A more plausible story is that the two-player model is a reduced form of a *larger* model, in which interpretation is done by a *third party* whose preferences are aligned with the sender's: An accomplice, a spokesperson or a captured media outlet. Such third parties provide selective data that illuminate the meaning of utterances by the agent they serve.

We could turn the interpreter into an actual third player, producing the following timeline. The sender moves first by choosing a message m. The interpreter moves after observing m (but not θ) and chooses D. This means that the interpretation strategy must be measurable with respect to m. Unlike the receiver, the interpreter has rational expectations. The conditional distribution σ over pairs (m, D) is induced by the combination of the message and interpretation strategies, and it is restricted to satisfy the conditional-independence property $D \perp \theta \mid m$. The receiver moves last, having observed the history (m, D), and he best-replies to the belief (2.1). If the sender and interpreter have common interests, the situation can be reduced to our two-player formulation, under a suitably defined solution concept for the three-player interaction. In Section 3 we will see that there is no loss of generality in imposing $D \perp \theta \mid m$ directly on the two-player model, lending support to this three-player interpretation of our model.

Thus, while we will adhere to the sender-receiver formal terminology, our model can be regarded as a description of a situation in which the sender and interpreter are separate entities who happen to share common interests.[3]

2.3 Analysis

We begin this section by presenting the rational-expectations benchmark for our model. In this case, which coincides with the "prosecutor" example in Kamenica and Gentzkow (2011), the probability of persuasion is maximized by the following message strategy (the dictionary component in the reporting strategy is redundant): In state Y, the sender plays $(1, ..., 1)$ with probability one, whereas in state N, he plays $(1, ..., 1)$ with probability

[3]In a previous version of the paper (Eliaz et al., 2018), we analyzed an extension in which the interpreter's preferences are aligned with the receiver's with some probability; the sender does not know the interpreter's type when choosing his message strategy.

$\pi/(1-\pi)$ and $(0,...,0)$ with the remaining probability. When the receiver gets the message $(0,...,0)$, he infers that $\theta = N$ for sure and takes the action n. When he receives the message $(1,...,1)$, his posterior is

$$\Pr(\theta = Y \mid m = (1,...,1)) = \frac{\pi \cdot 1}{\pi \cdot 1 + (1-\pi) \cdot \frac{\pi}{1-\pi}} = \frac{1}{2}$$

such that he is just willing to play y. Consequently, the overall probability of persuasion is

$$\pi + (1-\pi) \cdot \frac{\pi}{1-\pi} = 2\pi$$

This result crucially relies on the sender's ability to *commit* to a strategy ex-ante. Without the ability to commit, the probability of persuasion would be *zero* in any Nash equilibrium.

The following example demonstrates that in contrast to the rational-expectations benchmark, our model enables *full* persuasion as an equilibrium outcome.

Example 2: Full persuasion under $K = 3$ and $K = 4$

Let $K = 3$ and consider the following sender strategy (for convenience, we highlight the interpreted components in each report in boldface). In each state, he mixes uniformly over three reports:

State Y		State N	
m	D	m	D
111	$\{1\}$	**1**00	$\{1\}$
1**1**1	$\{2\}$	0**1**0	$\{2\}$
11**1**	$\{3\}$	00**1**	$\{3\}$

Notice that in state Y only one message is sent, but the sender randomizes the dictionary it is paired with. In contrast, in state N, three distinct messages are sent with three distinct dictionaries, where each dictionary interprets a pattern that also appears in state Y (namely, the component with the digit 1). For each of the six reports $(m, \{k\})$, the receiver's posterior belief $\widetilde{\Pr}(\theta = Y \mid m, \{k\})$ is

$$\frac{\pi \cdot \Pr(m_k = 1 \mid \theta = Y)}{\pi \cdot \Pr(m_k = 1 \mid \theta = Y) + (1-\pi) \cdot \Pr(m_k = 1 \mid \theta = N)} = \frac{\pi \cdot 1}{\pi \cdot 1 + (1-\pi) \cdot \frac{1}{3}} = \frac{3\pi}{1+2\pi}$$

The receiver weakly prefers playing y after each of these reports, as long as $\pi \geq \frac{1}{4}$.

If $K = 4$, the sender is able to achieve full persuasion for even smaller prior beliefs. He achieves this by using the following strategy, which in each state, uniformly randomizes over six reports:

State Y		State N	
m	D	m	D
$\mathbf{1}\mathbf{1}11$	$\{1,2\}$	$\mathbf{1}\mathbf{1}00$	$\{1,2\}$
$\mathbf{1}1\mathbf{1}1$	$\{1,3\}$	$\mathbf{1}0\mathbf{1}0$	$\{1,3\}$
$\mathbf{1}11\mathbf{1}$	$\{1,4\}$	$\mathbf{1}00\mathbf{1}$	$\{1,4\}$
$1\mathbf{1}\mathbf{1}1$	$\{2,3\}$	$0\mathbf{1}\mathbf{1}0$	$\{2,3\}$
$1\mathbf{1}1\mathbf{1}$	$\{2,4\}$	$0\mathbf{1}0\mathbf{1}$	$\{2,4\}$
$11\mathbf{1}\mathbf{1}$	$\{3,4\}$	$00\mathbf{1}\mathbf{1}$	$\{3,4\}$

For each of these twelve reports $(m, \{j,k\})$, the receiver's posterior belief $\widetilde{\Pr}(\theta = Y \mid m, \{j,k\})$ is

$$\frac{\pi \cdot \Pr(m_j = m_k = 1 \mid \theta = Y)}{\pi \cdot \Pr(m_j = m_k = 1 \mid \theta = Y) + (1-\pi) \cdot \Pr(m_j = m_k = 1 \mid \theta = N)} = \frac{\pi \cdot 1}{\pi \cdot 1 + (1-\pi) \cdot \frac{1}{6}}$$

The receiver weakly prefers playing y after each of these reports, as long as $\pi \geq \frac{1}{7}$.

This example illustrates a number of key points.

Non-rational expectations

The receiver reaches wrong beliefs as a result of the strategically chosen dictionaries. E.g., in the $K = 4$ case, although the reports $((1,1,1,1),\{2,3\})$ and $((0,1,1,0),\{2,3\})$ objectively reveal the state in which they are played, the receiver draws the same inference from both of them. The reason is that the two messages coincide on the second and third components, highlighted by the accompanying dictionary $\{2,3\}$.

Irrelevance of commitment

Since the sender achieves full persuasion, his strategy would also constitute an equilibrium in the *absence* of commitment. The reason is that the receiver plays y after any realized report, hence the sender has no incentive to deviate from any realization of his mixed strategy.

More (interpretation) can be less

The receiver is clearly harmed by selective interpretation: If the sender were compelled to interpret all message components, the problem would be effectively reduced to the rational-expectations benchmark. However, this effect is not monotone. Suppose that we made dictionaries even *more* selective by forcing them to be *singletons*. Then, in the $K = 4$ case, the sender would only be able to attain full persuasion when $\pi \geq \frac{1}{5}$, using a similar strategy to the one we presented for $K = 3$.

Dictionary-state independence

Our model assumes that the receiver cannot draw any inferences from D. Suppose he attempted such an inference - e.g. by acquiring data regarding the state-contingent distribution over dictionaries. Then, he would be unable to infer the state from D because its probability is identical in both states. One might argue that the receiver should still be "suspicious" of selective interpretations and discount their informational content. We devote Section 4 to this critique.

The sender's strategy satisfies another independence property: $D \perp \theta \mid m$. That is, given the realized message, the dictionary that accompanies it does not provide objective information about the state. This means that if the receiver had rational expectations, he could afford to draw inferences from m alone. The following lemma establishes that this property is not specific to the example.

Lemma 2.1. *The maximal probability of persuasion can be attained by a strategy that satisfies* $D \perp \theta \mid m$.

Proof. Consider an arbitrary sender strategy σ. Suppose that for a given message m there are two dictionaries D and D', such that both reports (m, D) and (m, D') are played with positive probability under σ. Suppose without loss of generality that the action induced by (m, D) is weakly more favorable to the sender (recall that the sender's preferences are state-independent). Consider a deviation that replaces (m, D') with (m, D). Since the deviation does not change the message strategy, it does not affect the receiver's reaction to any report $(m'', D'') \neq (m, D)$; and by increasing the probability of (m, D), it weakly increases the probability of persuasion. It follows that without loss of generality, we can assume that under the sender's strategy, every realized report m is accompanied by a *single* dictionary D_m. In particular, this means that D is independent of θ conditional on m. □

This lemma substantiates the three-player interpretation of our model that was described at the end of Section 2, since a distinct interpreter would only be able to condition D on m.

Should a dictionary interpret multiple messages?

The sender's strategy in Example 2 has a notable feature: In every report (m, D) that is played in state N, the pattern that D highlights does not appear in any other message that is played in N. Compare this with the report $((1,0,1,0),\{1,3\})$ in Example 1. The dictionary $\{1,3\}$ highlights the pattern $(1,*,1,*)$, which also appears in *another* message, $(1,1,1,1)$, that is played in state N. It turns out that this feature of the sender's behavior in Example 1 is weakly sub-optimal. That is, when solving the sender's problem, we can restrict attention to strategies that satisfy the following property: for every persuasive report (m, D) that is played in state N, D highlights a pattern that does not appear in any other message sent in that state. This property will facilitate the proof of our main result.

Fix a sender's strategy σ. Let $\mathcal{B}_\sigma$ be the set of reports (m, D) that are played with positive probability in $\theta = N$ and persuade the receiver. That is,

$$\mathcal{B}_\sigma = \left\{ (m, D) \mid \sigma(m, D \mid \theta = N) > 0 \text{ and } \rho_\sigma(m, D) \geq \frac{1-\pi}{\pi} \right\}$$

Proposition 2.1. *For every sender strategy σ, there exists a strategy σ' with the following properties: (i) the probability that the receiver chooses y in each state is at least as high as under σ, and (ii) $m'_D \neq m_D$ for every pair of distinct reports $(m, D), (m', D') \in \mathcal{B}_{\sigma'}$.*

Our proof employs a two-stage algorithm. In the first stage, we replace "redundant dictionaries". We list the reports in $\mathcal{B}_\sigma$ according to an arbitrary ordering. Then, starting with the report (m, D) at the top of the ordering, we identify messages m' down the list such that $m'_D = m_D$. We then replace the dictionaries that accompany these messages with D. Setting aside the top report and all the reports that were subjected to this replacement, we continue in the same manner with the remaining reports. At the end of the algorithm's first stage, $\mathcal{B}_\sigma$ is partitioned such that each cell consists of reports (m, D) with the same D and m_D. In the second stage, we replace "redundant messages". We go *up* the list of reports and modify messages only, such that each cell in the above partition ends up consisting of a single report. (We may perform additional changes to the dictionaries that accompany messages in state Y, to ensure that probability that $a = y$ in this state does

not go down.)

Our subsequent analysis makes use of the following concept.

Definition 2.1. *For a given a strategy σ, a message m' is said to* ***justify*** *the report $(m, D) \in \mathcal{B}_\sigma$ if: (i) $\sigma(m' \mid \theta = Y) > 0$, and (ii) $m'_D = m_D$.*

In other words, what helps persuade the receiver to choose y when he gets the report (m, D) is that the pattern highlighted by D appears in some messages m' that are played with sufficient frequency in state Y.

Proposition 2.1 is particularly useful because it places restrictions on the family of reports that any given message can justify. This is captured by the following corollaries.

Corollary 2.1. *Let $(m, D), (m', D') \in \mathcal{B}_\sigma$. If there is a message m^* that justifies both (m, D) and (m', D'), then $D \not\subseteq D'$ and $D' \not\subseteq D$.*

Corollary 2.2. *The number of reports that any message justifies is at most $\binom{K}{\lfloor K/2 \rfloor}$.*

Corollary 2.1 says that the set of dictionaries that appear in reports that are justified by a given message m^* constitutes an *anti-chain* - i.e., no dictionary in this set contains another. Corollary 2.2 then invokes Sperner's Theorem. This fundamental result in extremal combinatorics states that the largest anti-chain over $\{1, 2, \ldots, K\}$ is the collection of all subsets of size $\lfloor K/2 \rfloor$.

We are now ready to state the main result of this section. The result makes use of the following notation, which will also serve us in later sections:

$$S = \binom{K}{\lfloor K/2 \rfloor}$$
$$\mathcal{B}^* = \left\{ (m, D) \mid m_k = \mathbf{1}(k \in D) \ ; \ |D| = \left\lfloor \frac{K}{2} \right\rfloor \right\}$$

Note that $|\mathcal{B}^*| = S$.

Theorem 2.1. *The maximal probability of persuasion is* $\min\{1, \pi(1+S)\}$. *It can be implemented by the following strategy:*

$$\sigma((1,\ldots,1), D \mid \theta = Y) = \frac{1}{S} \text{ for every } D \text{ for which } |D| = \left\lfloor \frac{K}{2} \right\rfloor$$
$$\sigma(m, D \mid \theta = N) = \min\{\frac{1}{S}, \frac{\pi}{1-\pi}\} \text{ for every } (m, D) \in \mathcal{B}^*$$
$$\sigma((0,\ldots,0), D \mid \theta = N) = \max\{0, \frac{1}{S} - \frac{\pi}{1-\pi}\} \text{ for every } D \text{ for which } |D| = \left\lfloor \frac{K}{2} \right\rfloor$$

Furthermore, when $\pi \geq 1/(1+S)$, *this strategy is time-consistent and attains full persuasion.*

The strategy that implements the maximal probability of persuasion generalizes Example 2. In state Y, the sender sends a single message, which we conveniently select to be $(1, ..., 1)$. Each of the components of this message can therefore be regarded as "good news". What happens in state N depends on the relation between the prior π and the number S, which depends on K. Suppose K is even, for the sake of the argument. If $\pi \geq 1/(1+S)$, the sender randomizes uniformly over $\mathcal{B}^*$, which is the set of all reports in which the message consists of an equal number of 1's ("good news") and 0's ("bad news"), and the dictionary interprets only the good news. If $\pi < 1/(1+S)$, each of these reports is played with probability $1/S$, and the remaining probability is allocated to the message $(0, ..., 0)$ - i.e. all "bad news".

Unlike the case of the "mixed" messages in $\mathcal{B}^*$, there is considerable freedom in selecting the dictionaries that accompany the "pure" messages $(1, ..., 1)$ and $(0, ..., 0)$. Our construction has the property that $(\sigma(D \mid m = (1, ..., 1))$ and $(\sigma(D \mid m = (0, ..., 0))$ are both the same as the distribution over D conditional on $\mathcal{B}^*$. Consequently, the strategy satisfies the independence property $D \perp \theta$ (on top of the property $D \perp \theta \mid m$ that was established by Lemma 2.1). Thus, even if the receiver attempted to draw inferences from D, he would be unable to learn anything about θ from the realization of D itself.

As to the question of how large dictionaries should be (discussed in the context of Example 2), note that the sender's optimal strategy makes use of dictionaries that interprets exactly *half* of the message components.

Let us examine the receiver's reaction to various realized reports under the sender's strategy. When he confronts the message $(0, ..., 0)$, each of the dictionaries that accompany it interprets some "bad news", and the receiver learns that $\theta = N$ for sure. In contrast, every other realization of (m, D) satisfies $m_k = 1$ for all $k \in D$. The receiver thus learns that the

probability of m_D conditional on $\theta = Y$ is one, while the probability of m_D conditional on $\theta = N$ is $\min\{1/S, \pi/(1-\pi)\}$. The receiver's subjective likelihood ratio of (m, D) is

$$\rho_\sigma(m, D) = \frac{1}{\min\{\frac{1}{S}, \frac{\pi}{1-\pi}\}}$$

which is, by definition, weakly above $(1-\pi)/\pi$ and therefore persuasive.

A receiver with rational expectations would realize that the "mixed" messages in $\mathcal{B}^*$ only occur in state N. However, our receiver can only draw inferences from message components that the sender interprets for him. Since the sender only interprets persuasive patterns, he manages to convey a false sense that the mixed message is actually good news. Moreover, as K gets large, each $(m, D) \in \mathcal{B}^*$ identifies a distinct pattern that becomes increasingly rare in state N while occurring with probability one in state Y. Therefore, even when π is quite small and even if $\mathcal{B}^*$ is played with high probability in state N, the receiver will be persuaded by the reports in $\mathcal{B}^*$.

When $\pi \geq 1/(1+S)$, the sender can attain full persuasion. This means that the sender's strategy is *time-consistent*: Since the receiver plays $a = y$ after every report, the sender would not want to deviate from any realized report even if he could. In other words, the assumption that the sender has commitment power is not required in this range of parameters.

Theorem 2.1 assumes an unrestricted domain of feasible dictionaries. The proof of Theorem 2.1 makes the result easily extendible to restricted domains.

Remark 2.1. *Let $\mathcal{D}$ be the set of feasible dictionaries. Let $\mathcal{D}^* \subseteq \mathcal{D}$ be an anti-chain, such that every $\mathcal{D}' \subseteq \mathcal{D}$ with $|\mathcal{D}'| > |\mathcal{D}^*|$ is not an anti-chain. Then, the maximal probability of persuasion is* $\min\{1, \pi(1 + |\mathcal{D}^*|)\}$.

In particular, when the feasible set of dictionaries is the set of all *singletons*, the maximal probability of persuasion is $\max\{1, \pi(1+K)\}$. This suggests that if the sender were free to determine the dimensionality of the message space, he could trivially attain full persuasion with singleton dictionaries. However, K should be interpreted as an exogenous constraint: there is a *limited* set of variables about which statistical data is available. For instance, if message components correspond to non-verbal aspects of the sender's communication, only few of those aspects are typically documented (it is unlikely to have data about the sender's pupil dilation, blood pressure or EEG measurements). Similarly, if the sender is a political party and message components correspond to different party members, only the

messages of a few senior members are likely to be documented.

2.4 Suspicion of Selective Interpretations

In our discussion of Theorem 2.1, we raised the concern that the receiver may try to infer the state from the dictionary the sender provides. The sender strategy we presented in the theorem's statement addressed this concern, in the sense that it satisfied the independence property $D \perp \theta$. However, one may argue that even this feature would not quell the receiver's suspicion regarding the *selectiveness* of the provided dictionary - i.e., some message components are not interpreted. The receiver may view the mere neglect of message components as a signal that the state is N (even though the state-contingent distribution over dictionaries offers no basis for this suspicion).

While intuitive, this argument is actually unconventional. The receiver draws a correct Bayesian inference from the message components for which he gets data. In the absence of additional data on how dictionaries and messages are jointly distributed, there is nothing to guide the receiver on how to modify this Bayesian posterior. Any assertion that he should ignore his available data and conclude that the state must be N simply because he was given selective data by a strategic sender is merely an *additional assumption*. By the same token, one could argue that in the partially informative "interval equilibria" in Crawford and Sobel (1982), the receiver should ignore his statistical knowledge of the sender's behavior and trust *nothing* the sender says simply because he is known to lie or withhold information.[4]

This methodological discussion notwithstanding, we now address the possibility that receivers may be suspicious of selective interpretations by proposing two notions of robustness to this suspicion. In both cases, we show that full persuasion is attainable for a large range of parameters π, K, albeit smaller than in Theorem 2.1.

2.4.1 Benevolent Selectiveness

Even if the sender's interests were fully aligned with the receiver's, it would be reasonable for him to refrain from interpreting *all* message components and provide a selective dictio-

[4]If we interpret the sender's strategy as recommending an action or communicating the interval to which the state belongs, this is a case of witholding information. If we interpret his strategy as some mixture over states that belong to the interval, then his message misrepresents the state with probability one.

nary. To see why, let $K = 2$ and suppose that the message strategy is as follows: $m = (1, 1)$ with certainty in state Y, whereas $m = (0, 0)$ and $m = (1, 1)$ with equal probability in state N. Because m_1 and m_2 are fully correlated, the small dictionary $\{1\}$ induces the same receiver beliefs as the full dictionary $\{1, 2\}$. If the smaller dictionary is less costly to provide, a benevolent sender would use it (recall that D must be non-empty). In this case, the receiver would not be suspicious of the sender simply for providing a small dictionary.

To capture this idea, we modify our model by introducing an intrinsic preference for smaller dictionaries. Specifically, we assume that the sender has lexicographic preferences. His primary criterion is to maximize the probability that the receiver plays y. However, if he can induce the same receiver behavior with two alternative dictionaries D and D' such that $|D'| < |D|$, he prefers D' to D. In addition, we impose a refinement of the set of permissible sender strategies, which is based on a hypothetical *benevolent* sender. Such a sender has lexicographic preferences, too: His primary criterion is to maximize the receiver's payoff; his secondary criterion is to minimize $|D|$. Refer to this *hypothetical* sender as type H; whereas the *actual* sender will be referred to as type A.

Definition 2.2. *The strategy* $(\sigma(m, D \mid \theta))$ *is* ***robust*** *if it satisfies the following properties:*
(i) $D \perp \theta$ *and* $D \perp \theta \mid m$.
(ii) Given $(\sigma(m \mid \theta))$*, the interpretation strategy* $(\sigma(D \mid m))$ *prescribes, for each* m*, lexicographically optimal dictionaries for a type-A sender.*
(iii) Given $(\sigma(m \mid \theta))$*, there is an interpretation strategy* $(\sigma'(D \mid m))$ *that prescribes, for each* m*, lexicographically optimal dictionaries for a type-H sender, such that* $\sigma'(D) \equiv \sigma(D)$.

Condition (i) imposes the independence requirements we have already encountered in Section 3. Condition (ii) was redundant in Section 3 because we focused on optimal sender strategies anyhow. Here, it also means that the sender always uses the smallest dictionary that attains a given outcome.

As to condition (iii), our motivation is the following. Throughout the paper, we have assumed that the receiver lacks any data about the distribution of D. However, imagine now that the receiver has access to an independent dataset that enables him to learn the marginal distribution of dictionaries. (By condition (i), this is the same as learning the distribution of D at each state.) He can therefore see that the use of selective dictionaries is not a fluke, but an event that occurs with positive frequency. Condition (iii) requires further that if the dictionaries were chosen by a benevolent sender of type H, their marginal

distribution could be the same. From this point of view, the receiver is less likely to be suspicious of selective interpretations, because he can reconcile their observed statistical pattern with the existence of a benevolent interpreter having a lexicographically secondary preference for small dictionaries.

In what follows, we conveniently assume that the receiver always breaks ties in favor of a type-A sender.

Proposition 2.2. *Full persuasion is attainable with a robust strategy if and only if* $\pi \geq 1/(1+K)$.

Thus, requiring the sender's strategy to be robust in the sense of Definition 2.2 restricts his ability to attain full persuasion, because it effectively eliminates the use of non-singleton dictionaries. Example 2 in Section 3 illustrates a robust strategy that achieves full persuasion for $K = 3$.

2.4.2 Full-Coverage Dictionaries

In this subsection we use a different line of attack to address the selective-interpretation problem. Here, we assume that the sender is obliged to present statistical data about *all* message components. However, he is allowed to present the data in two separate chunks. As before, a dictionary is represented by a non-empty subset $D \subseteq \{1, ..., K\}$. Yet this now means that the sender provides *two* datasets, formalized as two collections of conditional probabilities: $(\Pr(m_D \mid \theta))$ as well as the $(\Pr(m_{D^c} \mid \theta))$, where $D^c = \{1, ..., K\} \backslash D$. We refer to this form of data provision as *full-coverage dictionaries*.

How does the receiver extrapolate a belief from the two datasets? We make the mild assumption that his subjective belief $\widetilde{\Pr}(m, D \mid \theta)$ satisfies

$$\Pr(m_D \mid \theta) \cdot \Pr(m_{D^c} \mid \theta) \leq \widetilde{\Pr}(m, D \mid \theta) \leq \max\{\Pr(m_D \mid \theta), \Pr(m_{D^c} \mid \theta)\} \tag{2.4}$$

The upper bound given by the R.H.S reflects an assumption that m_{D^c} is uninformative of θ given m_D, or vice versa - i.e., the two parts of m are perfectly correlated given the state. The lower bound given by the L.H.S reflects an assumption that these two parts are independent conditional on the state.

Proposition 2.3. *Let $K > 2$. Then, the sender can attain full persuasion with full-coverage dictionaries whenever*

$$\pi \geq \frac{4}{4+S}$$

Thus, although the sender is forced to provide data about *all* message components, his ability to present the data in two "installments" enables him to attain full persuasion for a large range of parameters. Moreover, the strategy we construct in the proof satisfies the familiar independence properties $D \perp \theta$ and $D \perp \theta \mid m$. Finally, the result relies on the relatively weak condition (2.4) on how the receiver extrapolates a belief from the two separate datasets he receives. Note that Proposition 2.3 only provides a sufficient condition for full persuasion. Finding a necessary condition is an open problem.

To illustrate the basic idea of the construction, let $K = 4$. In state Y, there is perfect correlation among all message components. The objective correlation is weaker in state N. Specifically, only the messages $(1,1,1,1)$ and $(0,0,0,0)$ are played in Y, whereas all messages containing exactly two 1's are played in N. Thus, patterns like $(*,1,1,*)$ or $(0,*,*,0)$ are considerably more likely in Y than in N. By accompanying the message $(0,1,1,0)$ with two datasets that separately highlight these two patterns, the sender can manipulate the receiver's likelihood ratio.

This subsection also illustrates that the form a dictionary can affect the sender's ability to persuade the receiver. This reinforces a point we made in Section 2: Our concept of "selective interpretation" is richer than what the "selective message redaction" metaphor might suggest.

2.5 Richer Dictionaries

In this section we follow up on the final paragraph of the previous section. So far, we have assumed that dictionaries provide data about the joint distribution of a collection of message components conditional on θ. However, statistical data can involve other combinations of marginal and conditional distributions, with implications for the sender's ability to persuade the receiver.

Example 3

Let $K = 2$. Let p denote the joint distribution over (θ, m) that is induced by the prior over θ and the sender's strategy. There are three feasible dictionaries: D_1 gives access to

the conditional distribution ($p(m_1 \mid \theta)$); D_2 gives access to the conditional distribution ($p(m_2 \mid \theta)$); and D_3 gives access to the marginal distribution ($p(m_1)$) *as well as* the conditional distribution ($p(m_2 \mid \theta, m_1)$). It does not contain data about how m_1 varies with θ.

The dictionaries D_1 and D_2 are familiar from Section 2; we apply the same belief-formation rule (2.1) for the receiver as in Section 2. However, D_3 is different because it provides *two* datasets. We assume that the receiver extrapolates a belief using the *maximum entropy principle* - i.e., his belief over (θ, m_1, m_2) maximizes (Shannon) entropy subject to the constraint that it is consistent with the marginal and conditional distributions he has learned. This principle has a rich tradition in AI (dating back to Jaynes (1957)). Spiegler (2020) has recently applied it in a similar context of games with players who extrapolate a belief from partial data. In the model of Section 4.2, the principle induces the L.H.S of (2.4). In the present context, the receiver's subjective distribution over messages conditional on the state, given D_3, is $\widetilde{\Pr}(m_1, m_2 \mid \theta) = p(m_1)p(m_2 \mid \theta, m_1)$.

Consider the following sender strategy:

State Y			State N		
m	D	$\Pr(m, D \mid Y)$	m	D	$\Pr(m, D \mid Y)$
$(1,1)$	D_3	ε	$(1,1)$	D_3	α
$(0,0)$	D_2	$1-\varepsilon$	$(1,0)$	D_2	β
			$(0,1)$	D_1	$1-\alpha-\beta$

We now show that for every $\pi > \frac{1}{10}\left(5-\sqrt{5}\right)$, there exist $\alpha, \beta, \varepsilon \in (0,1)$ such that the sender attains full persuasion with the above strategy.

Let us calculate the receiver's likelihood ratio for each report. Consider the report $((1,1), D_3)$. Our definition of the receiver's posterior belief given the dictionary D_3 implies the following likelihood ratio:

$$\frac{p(m_1=1)p(m_2=1 \mid \theta=Y, m_1=1)}{p(m_1=1)p(m_2=1 \mid \theta=N, m_1=1)} = \frac{1}{\frac{\alpha}{\alpha+\beta}} = \frac{\alpha+\beta}{\alpha}$$

Next, consider the reports $((0,0), D_2)$ and $((1,0), D_2)$. Since D_2 only interprets m_2, both reports induce the same subjective likelihood ratio:

$$\frac{p(m_2=0 \mid \theta=Y)}{p(m_2=0 \mid \theta=N)} = \frac{1-\varepsilon}{\beta}$$

Finally, consider the report $(0, D_1)$. Since D_1 only interprets m_1, this report induces the subjective likelihood ratio

$$\frac{p(m_1 = 0 \mid \theta = Y)}{p(m_1 = 0 \mid \theta = N)} = \frac{1 - \varepsilon}{1 - \alpha - \beta}$$

In order to attain full persuasion, the three likelihood ratios must all be weakly greater than $(1 - \pi)/\pi$. A straightforward calculation establishes that whenever $\pi > \frac{1}{10}\left(5 - \sqrt{5}\right)$, we can find $\alpha, \beta, \varepsilon$ that will satisfy these three inequalities. In particular, ε will be arbitrarily small. $\square$

Compare this finding with the result of Section 3. Given our original specification of dictionaries, the sender can attain full persuasion if and only if $\pi \geq \frac{1}{3}$. This is *higher* than the threshold we obtained in Example 3. The general problem of optimal persuasion under the broader definition of dictionaries as collections of marginal and conditional distributions remains open.

2.6 An Adversarial Sender

In this section we revisit the basic model of Section 2 and modify the sender's preferences, such that the sender-receiver interaction becomes a zero-sum game: In state Y (N), the sender's payoff is 1 if the receiver plays n (y) and -1 if he plays y (n). Rescale the receiver's payoff function to be minus the sender's payoff. In what follows, we assume that the receiver always breaks ties in the sender's favor.

Consider the rational-expectations benchmark in this case. On one hand, the receiver can guarantee an expected payoff of at least $\pi \cdot (-1) + (1 - \pi) \cdot 1 = 1 - 2\pi > 0$ by always playing n. On the other hand, the sender can force this expected payoff on the receiver by sending the same report in all states. Therefore, by the Minimax Theorem, the sender's equilibrium payoff in the rational-expectations benchmark is exactly $2\pi - 1 < 0$. In contrast, the following result establishes that in our model, the sender can attain the maximal possible payoff of 1 under the same condition as in Theorem 2.1, whenever $K \geq 3$.

Proposition 2.4. *Let $K \geq 3$. Then, whenever $\pi \geq 1/(1 + S)$, there is a strategy for the sender that induces a payoff of 1 with certainty.*

Proof. Construct the following strategy. Let $m_k \in \{0,1\}$ for every k. In state Y, the sender plays $m^* = (1,1,...,1)$ with probability one and accompanies this message with the dictionary $D = \{k\}$ for some arbitrary k. In state N, the sender assigns probability $(1-\gamma)/S$ to every (m,D) satisfying $m_k = 1$ for exactly $\lfloor K/2 \rfloor$ components k and $D = \{k \mid m_k = 1\}$, where γ is selected to be the unique solution of the equation

$$\frac{1}{\gamma + \frac{1}{S}(1-\gamma)} = \frac{1-\pi}{\pi}$$

The sender assigns the remaining probability γ to the message m^* and accompanies it with an arbitrary dictionary of size $\lfloor K/2 \rfloor$. This is a feasible strategy whenever $\gamma \in [0,1]$ or equivalently $\pi \in \left[1/(1+S), \frac{1}{2}\right]$.

By construction, $\rho(m,D) = (1-\pi)/\pi$ for every (m,D) that is played in state N, whereas

$$\rho(m^*, \{k\}) = \frac{1}{\gamma + \frac{1}{2}(1-\gamma)} < \frac{1-\pi}{\pi}$$

As a result, the receiver plays y in state N and n in state Y, generating a payoff of 1 for the sender. □

Thus, strategic interpretation can attain the sender's first-best even under maximal conflict of interests with the receiver. As in Section 3, this means that the commitment assumption is unnecessary.

However, the strategy we employed in the proof of this result violates two independence properties that we emphasized in Section 3: $D \perp \theta$ and $D \perp \theta \mid m$. Let us now see how to fix this limitation when $K \geq 3$ and $\pi \geq 1/K$. As before, $m_k \in \{0,1\}$ for every k. Let e_k denote the message m for which $m_k = 1$ and $m_l = 0$ for all $l \neq k$. For every m, let $-m$ denote the message m' for which $m'_k = 1 - m_k$ for every k. Now consider the following sender strategy. In state Y, he randomizes uniformly over all (m,D) such that $m = -e_k$ and $D = \{k\}$ for some $k = 1,...,K$. In state N, he randomizes uniformly over all (m,D) for which $m = e_k$ and $D = \{k\}$ for some $k = 1,...K$. It is easy to verify that $\rho(m,D) \geq (1-\pi)/\pi$ for every (m,D) that is played in N, while $\rho(m,D) \leq (1-\pi)/\pi$ for every (m,D) that is played in Y, as long as $\pi \geq 1/K$.

The following result expands the set of parameters for which the sender's first-best is attainable by a strategy that satisfies the two desiderata, making use of a more elaborate strategy.

Proposition 2.5. *Let $K = 2L$ for some integer $L > 1$. Then, there is a strategy that satisfies $D \perp \theta \mid m$ and $D \perp \theta$ and attains the sender's first-best whenever*

$$\pi \geq \frac{1}{1 + \binom{L}{\lfloor L/2 \rfloor}}$$

This result provides a sufficient condition for attaining the sender's first-best with a strategy that satisfies the two desiderata. The following table illustrates the strategy for $K = 4$ (the strategy induces the sender-optimal action in each state, as long as $\pi \geq \frac{1}{3}$):

State Y			State N		
m	D	$\Pr(m, D \mid Y)$	m	D	$\Pr(m, D \mid Y)$
0011	$\{1\}$	0.25	1000	$\{1\}$	0.25
0011	$\{2\}$	0.25	0100	$\{2\}$	0.25
1100	$\{3\}$	0.25	0010	$\{3\}$	0.25
1100	$\{4\}$	0.25	0001	$\{4\}$	0.25

Finding a tight necessary condition remains an open problem.

2.7 Related Literature

Our paper joins a small literature on strategic communication that departs from the standard paradigm of rational expectations under a common prior. Levy et al. (2018) study a sender-receiver model in which the receiver exhibits "correlation neglect". Specifically, the sender submits multiple simultaneous signals and the receiver erroneously treats them as being conditionally independent. This belief distortion is related to the model of Section 4.2. In that variant on our basic model, the receiver does not learn the state-contingent correlation between m_D and m_{D^c}. We allowed the receiver to hold a variety of beliefs regarding this correlation, including the possibility that they are conditionally independent, as in Levy et al. (2018). The reason that unlike Levy et al. (2018), the sender in our model can attain full persuasion is that he can tailor the data to the submitted message.

Patil and Salant (2020) consider a receiver (a statistician) who estimates a parameter based on a random sample whose size is strategically determined by an informed sender. As in our model, the receiver has boundedly rational expectations in the sense that he makes no inferences from the sample size he gets. Schwartzstein and Sunderam (2019) examine a persuasion game in which both parties observe a signal that is drawn from

a state-dependent distribution. The receiver's non-rational expectations are captured by the assumption that the sender knows the signal distribution, while the receiver believes in whatever signal distribution the sender reports. Galperti (2019) analyses a model of persuasion with non-common priors, where the sender can influence the receiver's prior belief. In particular, when the receiver observes a message that has zero probability according to his prior, he abandons it in favor of a new belief. We, on the other hand, maintain the common prior assumption but allow the sender to strategically determine the receiver's understanding of the equilibrium distribution.[5]

Our basic model of dictionaries and how the receiver reacts to them is closely related to the concept of analogy-based expectations equilibrium (ABEE) due to Jehiel (2005). According to this concept, players form coarse beliefs that are measurable with respect to an "analogy partition" of the possible states of the world. Our basic notion of a dictionary D as a subset of components of multi-dimensional messages corresponds to an analogy partition. A cell in the partition consists of all messages m with the same m_D. This version of the model can thus be viewed as an extensive game in which the sender chooses the message as well as the receiver's analogy partition (from a restricted domain of feasible partitions), and the solution concept is ABEE. (However, the variants of Sections 4.2 and 5 *cannot* be embedded in the ABEE framework.) This description raises a natural question: How well can the sender perform under an *unrestricted* domain of feasible analogy partitions? For the sake of brevity, we do not analyze this question here but in a separate note (Eliaz et al. (2019)).

Jehiel and Koessler (2008) modify the Crawford-Sobel model by assuming that the receiver bundles states into analogy classes according to an interval analogy partition. They show that certain analogy partitions give rise to ABEE with partial information transmission, even when the unique equilibrium under rational expectations is the babbling equilibrium. Hagenbach and Koessler (2019) analyze cheap-talk games where the sender aggregates the receiver's equilibrium strategy into analogy classes. In a similar vein, Mullainathan et al. (2008) study a cheap-talk game where the receiver uses a coarse analogy partition. In contrast to our model, the partitions in these papers are exogenous. Endogenous partitions arise in Jehiel (2011), where auction designer controls bidders' learning feedback regarding the distribution of past bids.

Glazer and Rubinstein, Glazer and Rubinstein (2012, 2014) study persuasion when the

[5]Independently of our paper, Salcedo (2019) considers a persuasion game with one sender who sends private messages to multiple *rational* receivers. The sender wishes to persuade at least m receivers in order to attain his objective. When $m = 1$, the sender's problem is essentially the same as the sender's problem in our model when he is restricted to singleton dictionaries.

sender is boundedly rational in the sense of having limited ability to misrepresent the state. They show that a rational receiver can construct intricate disclosure mechanisms that take advantage of this element of the sender's bounded rationality. Blume and Board (2013) and Giovannoni and Xiong (2019) study cheap talk when the receiver has uncertain ability to distinguish between distinct messages. In contrast to our framework, receivers in these papers have rational expectations and the sender is unable to influence their interpretative abilities.

Finally, Spiegler (2020) introduces a general framework for static games, in which the description of players' types includes "archival access", defined as selective data about correlations among the variables that constitute the state of the world. Dictionaries in our model are a form of archival access. Indeed, our model is an example of how to extend the formalism of Spiegler (2020) to sequential games. Our approach to modeling the receiver's partial understanding of the sender's strategy is also related to Glazer and Rubinstein (2019), where a "problem solver" has partial understanding of the equilibrium: He observes a summary statistic of the other players' strategies, and then best-replies to a uniform belief over all the strategy profiles that are consistent with this statistic.

2.8 Conclusion

Conventional models of strategic communication focus on the role of selective transmission of information. And yet, real-life communication also involves strategic *interpretation* of information. This paper formalized this aspect as selective provision of *statistical data* regarding the mapping from states to messages, under the assumption that this data is the sole basis for the receiver's inferences. In a pure persuasion model, we showed that strategic interpretation significantly enhances the sender's ability to persuade the receiver - to the point that *full* persuasion is sometimes possible, in sharp contrast to the standard rational-expectations benchmark.

From a broader perspective, the modeling innovation in this paper is the idea that one player can influence another player's understanding of equilibrium regularities, by affecting the statistical data regarding the equilibrium distribution that the latter player has at his disposal (his "archival access", to use the terminology of Spiegler (2020)) - just as in a standard extensive-form game, one player's information set can be determined by prior moves of other players. Exploring this idea outside the context of strategic communication is an interesting problem for future research.

References

Andreas Blume and Oliver Board. Language barriers. *Econometrica*, 81(2):781–812, 2013.

Vincent P Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, pages 1431–1451, 1982.

Kfir Eliaz, Ran Spiegler, and Heidi C. Thysen. Strategic interpretation. 2018.

Kfir Eliaz, Ran Spiegler, and Heidi C. Thysen. On persuasion with endogenous misspecified beliefs. 2019.

Simone Galperti. Persuasion: The art of changing worldviews. *American Economic Review*, 109(3):996–1031, 2019.

Francesco Giovannoni and Siyang Xiong. Communication under language barriers. *Journal of Economic Theory*, 180:274–303, 2019.

Jacob Glazer and Ariel Rubinstein. On optimal rules of persuasion. *Econometrica*, 72(6): 1715–1736, 2004.

Jacob Glazer and Ariel Rubinstein. A study in the pragmatics of persuasion: A game theoretical approach. *Theoretical Economics*, 1:395–410, 2006.

Jacob Glazer and Ariel Rubinstein. A model of persuasion with boundedly rational agents. *Journal of Political Economy*, 120(6):1057–1082, 2012.

Jacob Glazer and Ariel Rubinstein. Complex questionnaires. *Econometrica*, 82(4):1529–1541, 2014.

Jacob Glazer and Ariel Rubinstein. Coordinating with a "problem solver". *Management Science*, 65:2813–2819, 2019.

Jeanne Hagenbach and Frédéric Koessler. Cheap talk with coarse understanding. mimeo, 2019.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4): 620, 1957.

Philippe Jehiel. Analogy-based expectation equilibrium. *Journal of Economic theory*, 123 (2):81–104, 2005.

Philippe Jehiel. Manipulative auction design. *Theoretical economics*, 6(2):185–217, 2011.

Philippe Jehiel and Frédéric Koessler. Revisiting games of incomplete information with analogy-based expectations. *Games and Economic Behavior*, 62(2):533–557, 2008.

Ginger Zhe Jin, Michael Luca, and Daniel Martin. Is no news (perceived as) bad news? An Experimental Investigation of Information Disclosure, 2019.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Gilat Levy, Inés Moreno de Barreda, and Ronny Razin. Persuasion with correlation neglect: Media power via correlation of news content. 2018.

Sendhil Mullainathan, Joshua Schwartzstein, and Andrei Shleifer. Coarse thinking and persuasion. *The Quarterly journal of economics*, 123(2):577–619, 2008.

Sanket Patil and Yuval Salant. Persuading statisticians. 2020.

Bruno Salcedo. Persuading part of an audience. 2019.

Joshua Schwartzstein and Adi Sunderam. Using models to persuade. December 2019.

Ran Spiegler. Modeling players with random "data access". July 2020.

2.9 Appendix: Proofs

Proposition 2.1

Let σ be an optimal sender strategy. We now change it into a new strategy that satisfies the property in the statement of the proposition and does not lower the probability of persuasion. We proceed in two stages.

Stage 1. Construct a partition $\{T_1, ..., T_L\}$ of $\mathcal{B}_\sigma$ as follows. For every $l = 1, 2, ...$, select an arbitrary report $(m^l, D^l) \in \mathcal{B}_\sigma - \cup_{h<l} T_h$, and define

$$T_l = \{(m, D) \in \mathcal{B}_\sigma - \cup_{h<l} T_h \mid m_{D^l} = m^l_{D^l}\}$$

Modify σ as follows. For each $l = 1, ..., L$ and any $(m, D) \in T_l$ with $D \neq D^l$, shift the probability of (m, D), conditional on $\theta = N$, to the report (m, D^l). By the definition of $\mathcal{B}_\sigma$, both (m, D) and (m^l, D^l) persuade the receiver. Perform the following additional modification. By the definition of $\mathcal{B}_\sigma$, there must be a message m that justifies (m^l, D^l). That is, $m_{D^l} = m^l_{D^l}$, and there is a dictionary D such that (m, D) is played with positive probability in Y. If the receiver was persuaded by (m, D) in the original strategy, then shift the probability of every such (m, D) conditional on Y to (m, D^l). By construction, $m_{D^l} = m^l_{D^l}$. Therefore, (m, D^l) persuades the receiver. And since the deviation does not affect the distribution over messages conditional on any state, it does not change the receiver's response to any other realized report.

Stage 2. Start this stage by shifting the probability of any $(m, D^L) \in T_L$ conditional on $\theta = N$ to some report in T_L, denoted $(\tilde{m}^L, D^L)$. This effectively transforms T_L into a singleton $\{(\tilde{m}^L, D^L)\}$. By the construction of the first phase, every $(m, D^L) \in T_L$ satisfies $m_{D^L} = \tilde{m}^L_{D^L}$. Therefore, the deviation does not change the receiver's subjective likelihood ratio of $(\tilde{m}^L, D^L)$, such that he continues to be persuaded by this report. Moreover, by the construction of the first stage, for every $l < L$ and every $(m, D^l) \in T_l$, $m_{D^l} \neq \tilde{m}^L_{D^l}$. Therefore, the deviation does not affect the receiver's subjective likelihood ratio of $(m, D^l) \in T_l$ for all $l < L$.

Now suppose that for some $l < L$, we have transformed the cells $T_{l+1}, ..., T_L$ into singletons $\{(\tilde{m}^{l+1}, D^{l+1})\}, ..., \{(\tilde{m}^L, D^L)\}$ in such a manner. Suppose that there is some $(m, D^l) \in T_l$ such that $m_{D^h} \neq \tilde{m}^h_{D^h}$ for every $h > l$. Rename this report $(\tilde{m}^l, D^l)$, and shift the probability of any (m, D^l) conditional on N to $(\tilde{m}^l, D^l)$. Alternatively, suppose that for

every $(m, D^l) \in T_l$ there is some $h > l$ such that $m_{D^h} = \tilde{m}^h_{D^h}$. For any such (m, D^l), shift its probability conditional on N to one of the reports $(\tilde{m}^h, D^h)$ satisfying $\tilde{m}^h_{D^h} = m_{D^h}$. By the same logic as in the previous paragraph, the deviation in these two alternative cases does not affect the receiver's subjective likelihood ratio of any report.

At the end of the second stage, $\mathcal{B}_\sigma$ has been effectively transformed into the set $\{(\tilde{m}^1, D^1)\}, ..., \{(\tilde{m}^L, D^L$ which by construction satisfies the property in the lemma's statement.

In the next two corollaries, we restrict attention to sender strategies σ that satisfy Proposition 2.1.

Corollary 2.1

Assume, by contradiction, that there exist $(m, D), (m', D') \in \mathcal{B}_\sigma$ that are justified by a message m^* and $D \subseteq D'$. This means that $m^*_D = m_D$ and $m^*_{D'} = m'_{D'}$. Therefore, $m_{D\cap D'} = m^*_{D\cap D'} = m'_{D\cap D'}$. But $D \cap D' = D$, which implies that $m_D = m'_D$, in contradiction to Proposition 2.1.

Corollary 2.2

By Corollary 2.1, if m^* justifies two reports (m, D) and (m', D'), then D and D' do not contain one another. It follows that the set of all dictionaries that are part of reports justified by m^* constitutes an anti-chain - i.e. a collection of subsets of $\{1, ..., K\}$ that do not contain one another. By Sperner's Theorem, the maximal size of such a collection is S.

Theorem 2.1

To derive an upper bound on the probability of persuasion, we restrict attention to sender strategies σ that satisfy Proposition 2.1. We begin with a basic observation that simplifies notation and the construction of the sender's strategy that maximizes the probability of persuasion in the N event. Fix a sender' strategy.

Observation 2.1. *There is no loss of generality in restricting attention to strategies with the following property: If the reports $(m, D) \in \mathcal{B}_\sigma$ and $(m', D') \notin \mathcal{B}_\sigma$ are both realized with positive probability in the N state under σ, then $m'_D \neq m_D$.*

Proof. Assume the contrary - i.e. $m'_D = m_D$. Suppose the sender deviates to a strategy that replaces (m', D') with (m', D) in the N state, but otherwise coincides with σ. By definition of $\mathcal{B}_\sigma$, (m', D') does not persuade the receiver prior to the deviation. And since the deviation does not affect the distribution of messages conditional on any state, it does not change the response of the receiver to any report $(m'', D'') \neq (m', D')$. Therefore, the deviation weakly raises the probability of persuasion. □

Henceforth, we will restrict attention to strategies that satisfy Observation 2.1. In addition, whenever we refer to a generic report in the N state, we mean a report in $\mathcal{B}_\sigma$.

Lemma 2.2. *Without loss of generality, $\rho_\sigma(m, D)$ is the same for all $(m, D) \in \mathcal{B}_\sigma$.*

Proof. Let $(\underline{m}, \underline{D})$ and $(\bar{m}, \bar{D})$ be two reports in $\mathcal{B}_\sigma$ such that $\rho_\sigma(\underline{m}, \underline{D}) \leq \rho_\sigma(m, D) \leq \rho_\sigma(\bar{m}, \bar{D})$ for each $(m, D) \in \mathcal{B}_\sigma$. Assume that $\rho_\sigma(\underline{m}, \underline{D}) < \rho_\sigma(\bar{m}, \bar{D})$. Suppose that the sender deviates from σ to a strategy $\hat{\sigma}$ that shifts a weight of $\varepsilon > 0$ from $(\underline{m}, \underline{D})$ to $(\bar{m}, \bar{D})$ in state N. By Proposition 2.1, $\bar{m}_{\underline{D}} \neq \underline{m}_{\underline{D}}$ and $\underline{m}_{\bar{D}} \neq \bar{m}_{\bar{D}}$. Therefore,

$$\begin{aligned} \rho_{\hat{\sigma}}(\underline{m}, \underline{D}) &= \frac{\sum_{m|m_{\underline{D}}=\underline{m}_{\underline{D}}} \sigma(m \mid \theta = Y)}{\sum_{m|m_{\underline{D}}=\underline{m}_{\underline{D}}} \sigma(m \mid \theta = N) - \varepsilon} > \rho_\sigma(\underline{m}, \underline{D}) \geq \frac{1-\pi}{\pi} \\ \rho_{\hat{\sigma}}(\bar{m}, \bar{D}) &= \frac{\sum_{m|m_{\bar{D}}=\bar{m}_{\bar{D}}} \sigma(m \mid \theta = Y)}{\sum_{m|m_{\bar{D}}=\bar{m}_{\bar{D}}} \sigma(m \mid \theta = N) + \varepsilon} < \rho_\sigma(\bar{m}, \bar{D}) \end{aligned} \tag{2.5}$$

By our initial assumption, $\rho_{\hat{\sigma}}(\underline{m}, \underline{D}) < \rho_{\hat{\sigma}}(\bar{m}, \bar{D})$ for sufficiently small ε. By (2.5), this implies that $\rho_{\hat{\sigma}}(\bar{m}, \bar{D}) > \frac{1-\pi}{\pi}$. By Proposition 2.1 $\rho_{\hat{\sigma}}(m, D) = \rho_\sigma(m, D)$ for every $(m, D) \in \mathcal{B}_\sigma - \{(\underline{m}, \underline{D}), (\bar{m}, \bar{D})\}$. Since the deviation does not involve reports outside $\mathcal{B}_\sigma$, it cannot alter the probability of persuading the receiver for messages outside of $\mathcal{B}_\sigma$. It follows that the deviation does not alter the probability of persuasion.

Therefore, we can assume without loss of generality that $\rho_\sigma(m, D)$ is the same for all $(m, D) \in \mathcal{B}_\sigma$. □

The remainder of the proof computes an upper bound on the probability of persuasion. Let σ be a sender strategy. Let $\mathcal{M}_Y = \{m \mid \sigma(m \mid \theta = Y) > 0\}$. Denote $I = |\mathcal{M}_Y|$. Let $\mathcal{C} = \{C_1, \cdots, C_L\}$ be a partition of $\mathcal{B}_\sigma$, where each cell C_l is defined by the (distinct) subset of messages $J(l) \subseteq \mathcal{M}_Y$ that justify every report in the cell. Therefore, $L \leq 2^I - 1$. For the final piece of notation we let $g(l) = \mid C_l \mid$ and $\beta(l) = \sum_{(m,D)\in C_l} \sigma(m, D \mid \theta = N)$.

Consider some $(m, D) \in C_l \subseteq \mathcal{B}_\sigma$ and a message $m' \in J(l)$. Since m' justifies (m, D), $m'_D = m_D$. By Proposition 2.1, there cannot be a dictionary D' such that $(m', D') \in \mathcal{B}_\sigma$.

It follows that for any $l = 1, ..., L$, the receiver's subjective likelihood ratio of a report $(m, D) \in C_l \subseteq \mathcal{B}_\sigma$ is

$$\frac{\sum_{m' \in J(l)} \sigma(m' \mid \theta = Y)}{\sigma(m, D \mid \theta = N)} \geq \frac{1-\pi}{\pi}. \tag{2.6}$$

From lemma 2.2 we have $\rho(m, D) = \rho(m', D')$ for every $(m, D), (m'D') \in \mathcal{B}_\sigma$. So in particular for every $(m, D), (m'D') \in C_l$ we have $\sigma(m, D) = \sigma(m', D') = \frac{\beta(l)}{g(l)}$. We can therefore rewrite inequality 2.6 as:

$$\frac{\sum_{m' \in J(l)} \sigma(m' \mid \theta = Y)}{\frac{\beta(l)}{g(l)}} \geq \frac{1-\pi}{\pi}, \tag{2.7}$$

Solving for $\beta(l)$ in (2.7) and summing over l give us

$$\begin{aligned}\sum_{l=1}^{L} \beta(l) &\leq \sum_{l=1}^{L} g(l) \sum_{m' \in J(l)} \left[\frac{\pi}{1-\pi}\sigma(m' \mid \theta = Y)\right] \\ &= \sum_{m' \in M^*} \left[\frac{\pi}{1-\pi}\sigma(m' \mid \theta = Y)\right] \sum_{l \in J^{-1}(m')} g(l)\end{aligned}$$

where the second equality follows from changing the order of summation. By definition, $\sum_{l \in J^{-1}(m')} g(l)$ is the number of reports that are justified by m'. By Corollary 2.2, this number is at most S. Therefore,

$$\begin{aligned}\sum_{l=1}^{L} \beta(l) &\leq \sum_{m' \in M^*} \left[\frac{\pi}{1-\pi}\sigma(m' \mid \theta = Y)\right] S \\ &= \frac{\pi}{1-\pi} S\end{aligned} \tag{2.8}$$

where the final equality follows since $\sum_{m' \in M^*} \sigma(m' \mid \theta = Y) = 1$. Since the receiver can at most be persuaded with probability one, the upper bound on the probability of persuasion in the N state is

$$\min\left\{\frac{\pi}{1-\pi} S, 1\right\}.$$

Verifying that the strategy described in the statement of Theorem 2.1 implements the upper bound is straightforward. This completes the proof.

Proposition 2.2

Sufficiency. Use the notation e_k for the binary K-vector for which $m_k = 1$ and $m_l = 0$ for all $l \neq k$. Consider the following strategy: When $\theta = Y$, play $m = (1, ..., 1)$ with probability one and randomize uniformly over all $D = \{k\}$, $k = 1, ..., K$. When $\theta = N$,

randomize uniformly over all reports $(m, D) = (e_k, \{k\})$, $k = 1, ..., K$. It is easy to see that $\rho(m, D) = K$ for every (m, D) in the support of this strategy. Therefore, when $\pi \geq 1/(1 + K)$, the receiver always plays $a = y$. Let us now verify that the strategy is robust. First, by construction, the distribution over D is state-independent, thus satisfying part (i) in the definition of robustness. Second, given the message strategy, a type-H interpreter can attain his first-best with the following interpretation strategy: When $m = (1, ..., 1)$, he mimics the given interpretation strategy; and when $m = e_k$, he plays $D = \{k + 1 \bmod K\}$, thus inducing $a = n$ with the smallest possible dictionary.

Necessity. Suppose that σ is a robust strategy that attains full persuasion. Let $\mathcal{D}$ denote the set of all non-singleton dictionaries that are played with positive probability under σ. The proof will proceed stepwise, after making the following preliminary observation.

Observation 2.2. *Fix a message strategy $(\sigma(m \mid \theta))$ and consider two dictionaries D, D' such that $|D| \neq |D'|$. Then, for any realized m, neither sender type is indifferent between D and D'.*

This follows immediately from the lexicographic preferences.

Step 1: $\Pr(\mathcal{D}) < 1$.

Assume the contrary - i.e., no singleton dictionary is played in equilibrium. Consider a message realization m for which $\Pr(\theta = Y \mid m) < \frac{1}{2}$ under σ. Since $\pi < \frac{1}{2}$, there must exist such m. By the full-persuasion assumption, any D for which $\sigma(D \mid m) > 0$ satisfies $\rho(m, D) \geq (1 - \pi)/\pi$. By condition (ii) in the definition of robustness, it must be the case that

$$\rho(m, \{k\}) < \frac{1 - \pi}{\pi} \tag{2.9}$$

for every $k = 1, ..., K$ - otherwise, the type-A sender would use a singleton dictionary at m. It follows from (2.9) that a type-H interpreter would necessarily prefer to use a singleton dictionary at m. By condition (iii) in the definition of robustness, singleton dictionaries must be played with positive probability under σ, a contradiction. □

Step 2: *Suppose $|D| = 1$ for some report (m, D) that is played with positive probability under σ. Then, $|D'| = 1$ for every (m', D') that is played with positive probability under σ, such that $m'_D = m_D$.*

Assume the contrary - i.e. there exist reports (m, D) and (m', D') that are played with positive probability under σ, such that $|D| = 1$, $|D'| > 1$ and $m'_D = m_D$. By definition,

$\rho(m', D) = \rho(m, D)$. Therefore, the realization (m', D') is inconsistent with condition (ii) in the definition of robustness. $\square$

By Observation 2.2, we can partition the set of equilibrium messages into two classes: M_0 is the set of messages that are accompanied by singleton dictionaries, whereas M_1 is the set of messages that are accompanied by non-singleton dictionaries. Recall that

$$\Pr(m_D \mid \theta) = \sum_{(m',D')|m'_D=m_D} \sigma(m', D' \mid \theta)$$

By Step 2, if $m \in M_0$, the R.H.S summation only covers reports (m', D') such that $m' \in M_0$. Furthermore, by condition (i) in the definition of robustness, $\Pr(M_0 \mid \theta = Y) = \Pr(M_0 \mid \theta = N) = \alpha$ under σ. By Step 1, $\alpha > 0$.

It follows that we can rewrite the joint distribution over (θ, m, D) that is induced by σ as a three-stage lottery. In the first stage, *before* θ is realized, the classes M_0 and M_1 are drawn with probability α and $1 - \alpha$, respectively. In the second stage, θ is realized, where $\theta = Y$ with probability π, independently of the lottery's first stage. Finally, (m, D) is realized conditional on θ, with the restriction that m must belong to the class that was realized in the first stage.

Therefore, in order for the receiver to play $a = y$ with probability one, it must be the case in particular that he plays $a = y$ with probability one conditional on the realization M_0 in the first stage of the three-stage lottery. But this can only hold if the condition for full persuasion given in Remark 2.1 for the case of singleton dictionaries. Therefore, it must be the case that $\pi \geq 1/(1 + K)$.

Proposition 2.3

Construct the following strategy for the sender.

Message strategy. In state Y, the sender randomizes uniformly between $m = (1, ...1)$ and $m = (0, ..., 0)$. In state N, he randomizes uniformly over the set of all messages m for which $m_k = 1$ for exactly $\lfloor K/2 \rfloor$ values of k.

Interpretation strategy. Every m that is played in state N is accompanied by $D = \{k \mid m_k = 1\}$. In state Y, the sender mixes uniformly over all sets D of size $\lfloor K/2 \rfloor$, independently of m.

By construction, $\Pr(m_D \mid \theta = Y) = \Pr(m_{D^c} \mid \theta = Y) = \frac{1}{2}$ and $\Pr(m_D \mid \theta = N) = \Pr(m_{D^c} \mid \theta = N) = 1/S$ for every (m, D) that is played. By (2.4), the receiver's likelihood ratio for any realized message m satisfies

$$\begin{aligned}\frac{\widetilde{\Pr}(m, D \mid \theta = Y)}{\widetilde{\Pr}(m, D \mid \theta = N)} &\geq \frac{\Pr(m_D \mid \theta = Y) \cdot \Pr(m_{D^c} \mid \theta = Y)}{\max\{\Pr(m_D \mid \theta = N), \Pr(m_{D^c} \mid \theta = N)\}} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{S}} = \frac{S}{4}\end{aligned}$$

The receiver will play $a = y$ whenever this expression is weakly above $(1 - \pi)/\pi$.

Proposition 2.5

Denote $S(L) = \binom{L}{\lfloor L/2 \rfloor}$. Construct the message strategy first. In state Y, randomize uniformly over two messages: m^1 satisfies $m^1_k = 1$ for all $k \leq L$ and $m^1_k = 0$ for all $k > L$; m^2 satisfies $m^2_k = 0$ for all $k \leq L$ and $m^2_k = 1$ for all $k > L$. In state N, assign probability $\frac{1}{2}S(L)$ to every message m such that $m_k = 1$ for $\lfloor L/2 \rfloor$ values of $k \in \{1, ..., L\}$, and $m_k = 0$ for all other k. Likewise, assign probability $\frac{1}{2}S(L)$ to every message m such that $m_k = 1$ for $\lfloor L/2 \rfloor$ values of $k \in \{L + 1, ..., 2L\}$, and $m_k = 0$ for all other k.

The conditional dictionary distribution is as follows. Conditional on any m that is played in state N, let $D = \{k \mid m_k = 1\}$ with certainty. Conditional on m^1, D is distributed uniformly over all subsets of $\{L + 1, ..., 2L\}$ of size $\lfloor L/2 \rfloor$. Finally, conditional on m^2, D is distributed uniformly over all subsets of $\{1, ..., L\}$ of size $\lfloor L/2 \rfloor$.

It is easy to verify that this strategy satisfies the two desiderata and induces the sender's first-best whenever $\pi \geq 1/(1 + S(L))$.

Chapter 3

Political Budget Cycles under a Flexible Election Regime

3.1 Introduction

Political business cycles have been suggested as a tool for politicians that allows them to signal their type to voters. This has been analysed for institutions in which the timing of elections are fixed: first by Nordhaus (1975) with naive voters and later by Rogoff (1990) with rational voters. However, in many countries the incumbent has the power to call for a snap election at will within the term length. This implies that the election timing is not fixed, but endogenously determined. It raises the possibility that politicians may have more than one signalling tool, and that these tools can interact.

I analyse a model where I assume that elections have to take place in every second period. In off-election periods, the politician leader (incumbent) has the power to call for a snap election, and thereby has two periods before the next election has to take place. In every period the incumbent chooses how much to invest in a public good that benefits the voters in the subsequent period. The level of public investment determined by the fiscal policy (the level of lump sum taxes or subsidies) also depends on the incumbent's ability, which follows a first order moving average process. The voters observe past periods' ability shocks, whether the election is a snap election as well as the fiscal policy choices. Based on these observations the voters make inferences about the incumbent's current period ability shock (which influences the incumbent ability to produce public investment goods in the future). When the incumbent is not in office she is an ordinary citizen and thus in the absence of re-election concerns the preferences of the voters and the incumbent are

aligned. However, the incumbent gets a utility boost from being in office. Therefore, the incumbent may have an incentive to distort her fiscal policy choices to signal her ability, and thereby increase her expected time in office.

In the benchmark of full information I show that the incumbent never calls for a snap election when her current ability shock is low. However, there is a range of parameter values for which the incumbent calls for a snap election whenever her current ability shock is high. This happens whenever the expected loss from having an election is outweighed by the expected gain from not being forced to have an election in the next period (where her re-election probability may be worse). An equilibrium with snap elections leads to a lower citizen utility in the long run compared to when election timing is fixed.

In the case of asymmetric information, I show that when the ego-rents are high enough fiscal policy distortion and snap elections are signalling substitutes and the welfare loss of the voters under the flexible election regime is mitigated compared to the full information benchmark. The basic intuition is that when an incumbent faces a low probability of re-election, she strictly prefers to postpone the election for another period as this ensures at least one additional period in office and is likely to improve her re-election probability. This implies that the low ability incumbent has a strictly higher outside utility in periods without elections compared to election periods. Therefore, she has greater incentives to mimic the incumbent with a high ability shock in the periods where she is forced to face an election. Whether there exists an equilibrium in which the flexible election regime improve the voters' welfare remains an open question. For the parameter values in which the incumbent faces the same probability of election in an equilibrium with snap election as under the fixed election regime there are two main forces. In an equilibrium with snap elections, there is less distortion of fiscal policies, but the overall distribution of the political leader's ability is worse. I show that when fiscal policies are distorted regardless of the previous period's ability shock then the utility loss from policy distortion is higher when the ability shock from the previous period is high. This implies that an incumbent with a low ability shock is relatively more likely to get re-elected when there are fiscal policy distortions in periods with snap elections. However, an incumbent with a low ability shock *always* faces a lower probability of re-election than an incumbent with a high ability shock and there does not exist any separating equilibria in which such an incumbent calls for a snap election.

The paper proceeds as follows. The next subsection reviews the related literature. Section 3.2 introduces the model and the timing of events. Section 3.2.7 formalises the equilibrium.

The full information equilibria are analysed in section 3.3. Proposition 3.1 characterises when there exists a full information equilibrium without any snap elections and provides a sufficient condition for the existence of a full information equilibrium in which the incumbent calls for a snap election when her current period ability shock is high. Proposition 3.2 shows that the fixed election timing is always better in the face of full information. Section 3.4 analyses the more realistic situation of asymmetric information. Proposition 3.3 establishes properties that hold in any separating equilibrium. Sufficient conditions for existence of equilibria with and without snap elections are provided in Proposition 3.4. Section 3.5 concludes.

3.1.1 Related literature

This paper falls into two strands of literature. First and foremost the literature on political budget cycles, which started with the seminal papers of Rogoff (1990) and Rogoff and Sibert (1988) in which the voters are rational. Nordhaus (1975) was a precusor to this - in which voters are naive about the politicians economic manipulations. Carlsen (1997) provided a version of Rogoff (1990)'s model to accomodate a negative relationship between incumbent ability and political business cycles. Lohnmann (2003) provides a model similar to Nordhaus (1975) that allows the voters to have rational expectations. Alesina and Perotti (1997), Drazen (2000), Smith (2004) and Dubois (2016) provide good overviews of the development of the literature of political cycles.

Most modern democracies allow the political agents some freedom to influence the election timing. Therefore, the empirical study of political business cycles are tightly linked to that of timing of election. There is however no universal consensus on the results. Some papers primarily find support for political budget cycles (such as Blais and Nadeau (1992), Alesina et al. (1993), Schultz (1995), Reid (1998), de los Angeles Gonzalez (2002) and Brender and Drazen (2005)), or cycles in inflation (see Grier (1989), Alesina and Roubini (1992) and Carlsen (2007)). Schneider (2010) found evidence for political policy cycles in the absence of the ability to manipulate the budget. While other papers find the timing of elections seems to be prevalent (see Ito and Park (1988), Ito (1990), Cargill and Hutchinson (1991), Alesina et al. (1993) in the case of Japan, Heckelman and Berument (1998), Schleiter and Tavits (2016)). Yet others find support for both of these mechanisms (Chowdhury (1993) and Palmer and Whitten (2000), who find that the policies manipulated by the incumbent depend on their political convictions). Williams (2013) finds a correlation between elections and international disputes. Schleiter and Tavits (2016) and Aaskoven

(2020) estimate that the use of an early election leads to a 5 percentage point increase in vote share and a 5% increase in election funds, respectively. Smith (2003) find that called for an unexpectedly early election leads to worse electoral outcomes for the incumbent. Ferris and Olmstead (2017) argue that there is a sufficiency gain from introducing a fixed term election in Canada.

Although election timing has received extensive attention in empirical literature, the same cannot be said about theoretical literature. Baron (1998) and Lupia and Strom (1995) analyse endogenous election timing in the light of coalition formation. Chappell and Peel (1979), Lachler (1982) and Ginsburgh and Michel (1983) extend Nordhaus's (1975) model to allow for endogenous election timing. Smith (2003) analyses a model in which politicians are better at forecasting the future. Thus, in this model, calling for an unexpectedly early election is a bad signal of future economic performance. Balke (1990) and Kayser (2005) analyse early election timing as an optimal stopping problem. In Kayser (2005) the incumbent as in Rogoff (1990) has the ability to manipulate the voters' learning through policy choices. However, the model does not take the equilibrium effect of the continuation value into account, when analysing the optimal election timing. Baleiras and Santos (2000) and Canes-Wrone and Park (2012) also provide versions of Rogoff (1990) that allow for early elections, but in a 2(3) period model.

3.2 Modelling Framework

This is an infinite period model, $T = \infty$.

3.2.1 Preferences of the Representative Voter

Electoral base consists of a large number of (ex-ante) identical voters. Each of the voters values the consumption of a private and a public good. The representative voter wants to maximise her expected utility, $E_t^P[\Gamma_t]$, where E_t^P denotes the expectation conditional on the public's information set at time t,

$$\Gamma_t := \sum_{s=t}^{\infty} [U(c_s) + V(k_s) + F(\eta_s)]\beta^{s-t}, \tag{3.1}$$

c is the voter's consumption of the private good and k is provision of the public investment good. Assume that c and k are both normal goods. U and V are assumed to satisfy the

usual Inada conditions [1], and $\lim_{k\to 0} V(k) = -\infty$. $\beta < 1$ is the common discount factor, η is a random popularity shock, which will be discussed in more details below, $F(x)$ is a function that equals $x/2$ when the incumbent is re-elected and $-x/2$ otherwise and the t subscripts denote time.

3.2.2 Technology

At the beginning of each period, all voters obtain y units of a non-storable good, which can either be consumed privately or used as an input in the production of the public investment good. The amount used in the production of the public investment good is given by the period t lump sum tax, τ_t. The remaining $y - \tau_t$ units are the voter's private consumption in period t, c_t.

The period t budget for public investment good provision is given (per capita) by

$$k_{t+1} = \tau_t + \epsilon_t,$$

where ϵ_t is the competency of the politician in charge in period t, and the production of the public goods take a period to become beneficial to the voter. At the end of period $t+1$, k_{t+1} perish.

3.2.3 Stochastic Structure

All voters are possible leaders, but they differ in their innate ability to produce public investment goods. For each voter i the innate ability evolves according to a MA(1) process:

$$\epsilon_t^i = \alpha_t^i + \alpha_{t-1}^i,$$

where each α_t^i is independent drawn from a Bernoulli distribution with $\rho = \Pr(\alpha_t^i = \alpha^H)$, and $\alpha^H > \alpha^L > 0$. The αs are drawn independently across time and voters. When there is no room for confusion the superscript is dropped for the incumbent.

Although I formally model citizens' ability to produce public investment goods as something that changes over time, my preferred interpretation is that it is the world rather than ability that is ever changing. Thus, if the incumbent is very good at dealing with current problems, then she might be mediocre (but not bad) at dealing with next year's problems.

[1] f satisfy the Inada conditions if $f : \mathbb{R} \to \mathbb{R}_+$ is a continuously differentiable function, such that (1) $f(0) = 0$, (2) f is strictly increasing and concave, (3) $\lim_{x\to 0} \partial f(x)/\partial x = \infty$, and $\lim_{x\to\infty} \partial f(x)/\partial x = 0$.

However, further out into the future there is no telling how useful the incumbent's skill will be.

Furthermore, before voting the voters receive a common popularity shock towards the incumbent η_t, where $\eta \sim U\left[-\frac{1}{2\Psi}, \frac{1}{2\Psi}\right]$ with η_t and η_s independent for all $t \neq s$. $\eta_t > 0$ indicates that the voters prefer the incumbent to the opponent above and beyond their ability to produce public investment good. The popularity shock covers how charismatic the voters find the incumbent relative to the opponent.

3.2.4 The Incumbent's Utility Function

The politician in charge is drawn from the pool of voters. Therefore, she values the private consumption good and the public investment good the same way any other voters do. In addition she receives an ego-rent of X per period in office, where $0 < X < \infty$. The utility of the incumbent is therefore given by

$$E_t^I[\Gamma_t] + \sum_{s=t}^{\infty} \beta^{s-t} X \pi_{t,s},$$

where E_t^I denotes the expectation conditional on the incumbent's information set at time t, Γ is given by (3.1), and $\pi_{t,s}$ is the incumbent's estimate of being in office at time s conditional on being in office at time t. The population of voters is taken to be sufficiently large such that the probability that a given voter is elected for office is minuscule. This is why the ego-rents are not featured in the utility of the representative voter and for the incumbent after she leaves office.

3.2.5 Structure of Elections

The country's constitution specifies the election structure. The maximal term length is two years. After which an election must be held. The voters can either re-elect the current incumbent or elect the opposition leader, who is randomly drawn from the pool of voters. Under the flexible election regime the incumbent has the power to call a snap election in non-election years. There are no term limits. $\theta_t \in \{0, 1\}$ denotes the number of periods before an election has to take place ($\theta_t = 1$ if there was an election in period $t-1$), and $e_t \in \{0, 1\}$ denotes whether or not an election will take place at the end of period t.

3.2.6 Information Structure and Timing of Events

At the beginning of period t, the incumbent observes her new ability shock α_t and the number of periods before the next election has to take place θ_t. She chooses the tax level τ_t and whether to call for a snap election if possible, e_t (if $\theta = 0$ then $e_t = 1$). The voters observe the public investment made in period $t-1$, k_{t-1}, the incumbent's ability shock from the previous period α_{t-1},[2] θ_t, e_t and τ_t, and utilise this information to draw inferences about the public investment k_{t+1} and the incumbent's current ability shock. These inferences will be confirmed in the next period, but not before. The incumbent therefore has a temporary information advantage over the voters. However, the voters' inferences are correct and in any separating equilibrium (which is the main focus of this paper) they can deduce the incumbent's private information.

The information structure here is plausible, when it is costly for the voters individually to closely monitor and evaluate the government's performance. When there is no uncertainty about α_{t-2}, then it is costless for the voters to infer α_{t-1} given $\theta_{t-1}, e_{t-1}, \tau_{t-1}$ and k_t. Thus, the assumption that the voters observe α_{t-1} directly only has bite in the first period after a new politician takes office.

The voters have no way of inferring the ability shock of the opponent leader α_t^O, where O superscripts denote the opponent. However, the distribution of α is common knowledge.

At the end of period t the voters observe the popularity shock η_t. The popularity shock encompasses issues the voters might care about that are orthogonal to the candidates ability to generate public investment goods, such as the latest gossip about one of the candidates.

If $e_t = 0$, the period ends after the popularity shock has been realised and the incumbent continues into period $t+1$ as the leading politician. If $e_t = 1$, then an election takes place after the popularity shock is realised. The representative voter decides which candidate goes into office in the next period. He votes for the candidate that maximises his expected utility. If $v = 1$ denotes a vote for the incumbent and $v = 0$ a vote for the opponent, then

$$v_t = \begin{cases} 1 & \text{if } E_t^P[\Gamma_{t+1}] \geq E_t^P[\Gamma_{t+1}^O] \\ 0 & \text{otherwise.} \end{cases}$$

[2]This assumption is made to ensure that there is no residual uncertainty about the previous ability shock, when the opposition leader takes office.

3.2.7 Markov Perfect Equilibrium

In this subsection, we define the markov perfect equilibrium. Before we define the equilibrium, we recap the timing and decisions within each period: (1) In the beginning of every period t nature draws $\alpha_t \in \{\alpha^L, \alpha^H\}$ as described above. (2) After observing α_t the incumbent chooses the tax level and whether to call for a snap election (when possible) $(\tau_t, e_t) \in A^I(\alpha_{t-1}, \alpha_t, \theta_t)$ where $A^I(\alpha_{t-1}, \alpha_t, 0) = [-\alpha_{t-1} - \alpha_t, y] \times \{1\}$, and $A^I(\alpha_{t-1}, \alpha_t, 1) = [-\alpha_{t-1} - \alpha_t, y] \times \{0, 1\}$. (3) Nature then draws the incumbent's popularity shock η_t. (4) Finally, the voters observe $(\alpha_{t-1}, \theta_t, \tau_t, e_t, \eta_t)$ before electing next period's incumbent. That is $v_t \in A^v$ where $A^v = \{0, 1\}$ ($A^v = \{1\}$) when $e_t = 1$ ($e_t = 0$). The elected politician becomes the incumbent in period $t + 1$ and $\theta_{t+1} = e_t$.

As the only variables from previous periods that restrict the agents' actions are α_{t-1} and θ_t, it is therefore natural in this setting to restrict attention to strategies that only depend on the history of play through the variables α_{t-1} and θ_t. Thus, we restrict attention to Markov Perfect Equilibria in pure strategies, with the state variables (α_{t-1}, θ_t). The incumbent's pure markov strategy is given $\sigma^I : \{\alpha^L, \alpha^H\}^2 \times \{0, 1\} \to A^I$, the representatives voter's belief about $\alpha_t = \alpha^H$ is given by a function $\hat{\rho} : \{\alpha^L, \alpha^H\} \times \{0, 1\} \times A^I \to [0, 1]$, and the voter's strategy is given by $v : [0, 1] \times \left[-\frac{1}{2\Psi}, \frac{1}{2\Psi}\right] \to A^v$.

Definition 3.1 (Equilibrium)**.** *The triple $(\sigma^I, \hat{\rho}, v)$ is a Markov Perfect Equilibrium in pure strategies (henceforth equilibrium) if (1) the incumbent chooses the fiscal policy and election timing (whenever possible) to maximise her expected continuation value:*

$$\sigma^I \in \arg \max_{(\tau,e)\in A^I} E_t^I [\Gamma_t] + \sum_{s=t}^{\infty} \beta^{s-t} X \pi_{t,s}(\hat{\rho}, e).$$

(2) when an election takes places the representative voter always elects the politician which gives him the highest expected value given the voter's belief:

$$v = \begin{cases} 1 & \text{if } \hat{\rho} E_t^P[\Gamma_{t+1} \mid \alpha_t = \alpha^H] + (1 - \hat{\rho}) E_t^P[\Gamma_{t+1} \mid \alpha_t = \alpha^L] \geq E_t^P[\Gamma_{t+1}^O] \text{ or } e = 0 \\ 0 & \text{otherwise,} \end{cases}$$

and (3) the voter's belief about the incumbent's current equilibrium ability shock $\hat{\rho}$ is given by Bayes rule whenever possible.

3.3 Full Information Case

Before analysing the effects of flexible election timing under asymmetric information, it is useful to consider the effects under full information, i.e. when α_t is observable to both the incumbent and the voters in period t. The main result of this Section is Proposition 3.1 that shows that snap elections can arise in equilibrium when $\alpha_t = \alpha^H$, and Proposition 3.2 that shows this can have detrimental effects for welfare.

When the voters observe α_t in period t, there is no uncertainty about the incumbent's ability. We therefore have $\hat{\rho} = \mathbb{1}_{\alpha_t = \alpha^H}$, and

$$v = \begin{cases} 1 & \text{if } E_t^P[\Gamma_{t+1} \mid \alpha_t] \geq E_t^P[\Gamma_{t+1}^O] \\ 0 & \text{otherwise,} \end{cases}$$

Similar to Grossman and Hart (1983) the incumbent's decisions in period t, can be viewed as a two-stage procedure: in the first stage the incumbent chooses the optimal fiscal policy choices given an (no) election takes place at the end of period t, $e_t = 1$ ($e_t = 0$). From this we derive the incumbent's indirect utility as a function of e_t. In the second stage, the incumbent then chooses $e_t \in \{0, 1\}$ to maximise her indirect utility.

Stage 1: When the voter directly observes α_t, the incumbent's pre-election policy choice cannot influence her re-election probability by manipulating the voter's posterior beliefs about her current ability shock. Thus, τ_t is independent of θ_t, and e_t. This observation in combination with the simple production technology and storage process, the incumbent's fiscal policy choices can be broken down into a series of static maximization problems:

$$\max_{c_t, \tau_t, k_{t+1}} U(c_t) + \beta V(k_{t+1}), \quad \forall t \geq \bar{t},$$

subject to $c_t = y - \tau_t$, $k_{t+1} = \tau_t + \epsilon_t$, $k_{t+1}, c_t \geq 0$ and $k_{\bar{t}} = \bar{k}$. By substituting c_t and k_{t+1} into the maximization problems we derive the following first order condition with respect to τ_t:

$$-U'(y - \tau_t) + V'(\tau_t + \epsilon_t) = 0.$$

As U and V both satisfy the Inada conditions the first order condition is both necessary and sufficient, and the solution is unique.[3] We denote by $x^{FI}(\epsilon)$ the optimal policy choice

[3] To see this, note that as $\epsilon \to y$, then $U'(y - \tau) \to \infty$, so $U'(y - \tau) > \beta V'(\tau + \epsilon)$. Similarly, when $\tau \to -\epsilon$, then $V'(\tau + \epsilon) \to \infty$, so $U'(y - \tau) < \beta V'(\tau + \epsilon)$.

of the variable x, when the incumbent's ability is ϵ. Let

$$W^{FI}(\epsilon) = U(y - \tau^{FI}(\epsilon)) + \beta V(\tau^{FI}(\epsilon) + \epsilon)$$

denote the indirect utility when the incumbent's ability is ϵ (henceforth citizen utility). It is easy to see that $W^{FI}(\epsilon)$ is increasing in ϵ. Furthermore, as both private consumption c and public investment good k are both normal goods, $c^{FI}(\epsilon)$ and $k^{FI}(\epsilon)$ are both increasing in ϵ and $\tau^{FI}(\epsilon)$ is decreasing.

Because ϵ follows a MA(1) process and the popularity shock is temporary, the voter's expected utility is the same for both candidates from period $t+2$ onwards. Thus, if $e_t = 1$, the incumbent is re-elected, $v_t = 1$, if and only if

$$\eta_t \geq E_t^P[W^{FI}(\epsilon_{t+1}^O)] - E_t^P[W^{FI}(\epsilon_{t+1})].$$

When the voter knows that $\alpha_t = \alpha^i$ $i = L, H$, then

$$E_t^P[W^{FI}(\epsilon_{t+1})] = \rho W^{FI}(\alpha^i + \alpha^H) + (1-\rho)W^{FI}(\alpha^i + \alpha^L),$$

which we denote Ω^i. Similarly, we denote

$$E_t^P[W^{FI}(\epsilon_{t+1}^O)] = \rho^2 W^{FI}(2\alpha^H) + 2\rho(1-\rho)W^{FI}(\alpha^L + \alpha^H) + (1-\rho)^2 W^{FI}(2\alpha^L)$$

by Ω^O. Clearly, $\Omega^H > \Omega^O > \Omega^L$. So in the absence of a popularity shock, the voter re-elects an incumbent if and only if the current period ability shock is high. With the popularity shock the probability of being re-elected is higher for an incumbent with a high current period ability shock, $\alpha_t = \alpha^H$. In particular, the representative voter will re-elect the incumbent with current ability shock $\alpha_t = \alpha^i$ if $\eta \geq \Omega^O - \Omega^i$.

Stage 2: When $\theta_t = 0$, there has to be an election at the end of period t, and hence the analysis ends here. However, when $\theta_t = 1$, the incumbent can call for a snap election. In the absence of any popularity shock an incumbent with $\alpha_t = \alpha^H$ ($\alpha_t = \alpha^L$) will clearly (not) call for a snap election as the incumbent in this case is ensured a win (loss). However, with the popularity shock the picture is less clear-cut. The decision depends both on the current and future probability of being re-elected. Let $\pi_{t,s}(\hat{\rho}, e_t)$ denote the probability that incumbent is in office in period s when e_t is the election status, and $\hat{\rho}$ is the public's

belief about $\alpha_t = \alpha^H$. When there is no risk of confusion we denote by

$$\pi^i = \pi_{t,t+1}(\mathbb{1}_{i=H}, 1) \text{ and } \pi^\rho = \pi_{t,t+1}(\rho, 1) = \rho\pi^H + (1-\rho)\pi^L \tag{3.2}$$

the probability of winning an election, when it is known that $\alpha_t = \alpha^i$, and when the ability of the incumbent is unknown, respectively.

A necessary condition for an equilibrium in which the incumbent with $\alpha_t = \alpha^i$ always (never) calls for a snap election is:

$$\begin{aligned} & \Omega^i\pi^i + \Omega^O(1-\pi^i) + \sum_{s=1}^{\infty} \beta^{s-1} X \pi_{t,t+s}(\mathbb{1}_{i=H}, 1) \\ \geq (\leq) \quad & \Omega^i + X + \sum_{s=2}^{\infty} \beta^{s-1} X \pi_{t+1,t+s}(\rho, 1). \end{aligned} \tag{3.3}$$

The main trade-off for the incumbent is whether to have the risk of an election today, and thereby be able to avoid an election in the next period, or avoid the risk today knowing that there will be an election in the following period.

To rule out uninteresting cases in which an incumbent with $\alpha_t = \alpha^L$ calls for a snap election because ego-rents are so low, that she prefers to have the opponent in office (as this would increase the citizen utility in the next period).

Assumption 3.1 (Strong office motive)**.** $\Omega^O - \Omega^L < X$

That is, even if the incumbent knows that she will be ill-prepared to deal with next year's challenges compared to the opponent, the ego-rents from being in office are high enough that she still wants to be in office. Although strong office motives ensure that the incumbent never calls for a snap election to get out of office, this does not shed light on when the incumbent calls for a snap election for opportunistic reasons. The next proposition addresses this question.

Proposition 3.1. *Given Assumption 3.1.*

1. *There exists no equilibrium in which the incumbent calls for a snap election when* $\alpha_t = \alpha^L$.

2. *There exists a full information equilibrium in which the incumbent never calls for a snap election if and only if*

$$\frac{\Omega^H - \Omega^O}{X}(1 - \pi^H) \geq \frac{2(1+\beta)\pi^H - (2+\beta)}{2 - \beta^2}.$$

3. *There exists a full information equilibrium in which the incumbent calls for a snap election whenever* $\alpha_t = \alpha^H$ *if*

$$\frac{\Omega^H - \Omega^O}{\bar{X}}(1 - \pi^H) \leq \pi^H \left(1 + \beta[\rho\pi^H + 1 - \rho]\right) - \left(1 + \beta\frac{1}{2}\right),$$

where $\bar{X} = \frac{2X}{2-\beta^2(1+\beta[\rho\pi^H+1-\rho])}$.

The proof is given in the next subsection. Having strong office motives deter the incumbent from calling a snap election as a means of retiring from office, as the incumbent with $\alpha_t = \alpha^L$ faces a low probability of being re-elected if an election were to take place in the current period. However, by postponing the election one period, there is a chance she will recover (the situation will change for something more suited for her skill set), and her expected re-election probability will therefore increase in the next period. When $\alpha_t = \alpha^H$, the incumbent is relatively well-equipped to deal with the challenges in the near future. Therefore, she faces a relatively high re-election probability if there is an election in the current period compared to the expected re-election probability in the next period. In order to take advantage of this, she will have to face the risk of election today. Proposition 3.1.(2) provides a condition that is both necessary and sufficient for the incumbent to be unwilling to take that risk in equilibrium. The RHS of the inequality in Proposition 3.1.(2) is increasing in the discount rate. Thus, this is more likely to hold when β is low. The absence of a pure strategy equilibrium with no snap elections does not imply that there is a pure strategy equilibrium with snap elections, as the incumbent's strategy influences the continuation value of winning an election. When the incumbent calls for a snap election whenever $\alpha_t = \alpha^H$, then the expected ego-rents from being in office cannot be written as a geometrical sum. However, it is possible to bound the expected rent from below. Proposition 3.1.(3) uses such a lower bound and provides a sufficient condition for when the incumbent is willing to face the risk of a snap election in order to utilise her high

re-election probability. This condition is akin to a situation in which there are only three periods and the incumbent with $\alpha_1 = \alpha^H$ prefers to have an election in the first period *and* an election in the second period only if $\alpha_2 = \alpha^H$ over having a single election in period 2 with the ego-rents properly adjusted. As

$$\frac{2\pi^H\left(1+\beta[\rho\pi^H+1-\rho]\right)-(2+\beta)}{2-\beta^2\left(1+\beta[\rho\pi^H+1-\rho]\right)}$$

is increasing in β, the condition in Proposition 3.1.(3) is more likely to hold when β is high.

As the incumbent only calls for a snap election when she is relatively well-equipped to deal with next year's challenges, snap elections decrease the voters' welfare. This is easy to see if we compare average citizen utility over a full election cycle as t goes to infinity.

Proposition 3.2. *As $t \to \infty$, the average citizen utility is strictly higher under the fixed election regime compared to the flexible election regime when snap elections take place.*

When there are no snap elections the distribution over the incumbent's ability does not converge, instead it is periodic with two periods. In any period immediately following an election the probability that the incumbent's ability shocks (α_{t-1}, α_t) are (α^H, α^H), (α^H, α^L), (α^L, α^H) and (α^L, α^L) equal $\rho^2\left(\pi^H + 1 - \pi^\rho\right)$, $\rho(1-\rho)\left(\pi^H + 1 - \pi^\rho\right)$, $\rho(1-\rho)\left(\pi^H + 1 - \pi^\rho\right)$ and $(1-\rho)^2\left(\pi^L + 1 - \pi^\rho\right)$, respectively; whereas in periods before an election the probabilities equal ρ^2, $\rho(1-\rho)$, $\rho(1-\rho)$ and $(1-\rho)^2$, respectively.

In an equilibrium with snap elections the distribution of the incumbent's ability shocks converge. The distribution is given by the next lemma. Let ν_t be the probability that the incumbent calls for an early election, δ_t be the probability that there is no election and ξ_t^i be the probability that $\alpha_t = \alpha^i$ and the incumbent is forced to have an election, for $i = L, H$. Finally, let $\tilde{x} = \lim_{t\to\infty} x_t$.

Lemma 3.1. *In an equilibrium in which the incumbent calls for a snap election when* $\alpha_t = \alpha^H$*, we have* $\tilde{\nu} = \frac{\rho}{2-\rho}$*,* $\tilde{\delta} = \frac{1-\rho}{2-\rho}$*,* $\tilde{\xi}^H = \frac{\rho(1-\rho)}{2-\rho}$ *and* $\tilde{\xi}^L = \frac{(1-\rho)^2}{2-\rho}$*.*

The incumbent's ability shocks (α_{t-1}, α_t) *equals*

- (α^H, α^H) *with probability* $\frac{\rho^2}{2(2-\rho)}\big(1+\rho+4\pi^H(1-\rho)\big)$*,*
- (α^H, α^L) *with probability* $\frac{\rho(1-\rho)}{2(2-\rho)}\big(1+\rho+4\pi^H(1-\rho)\big)$*,*
- (α^L, α^H) *with probability* $\frac{\rho(1-\rho)}{2(2-\rho)}\big(4-4\rho\pi^H+\rho\big)$ *and*
- (α^L, α^L) *with probability* $\frac{(1-\rho)^2}{2(2-\rho)}\big(4-4\rho\pi^H+\rho\big)$*.*

Using Lemma 3.1 we can calculate the average citizens' utility over an election cycle.

3.3.1 Proof of Proposition 3.1

First, we observe that we can write the probability of being in office in period $t+s$ as $\pi_{t,s}(\hat{\rho}, e_t) = \pi_{t,\bar{t}}(\hat{\rho}, e_t)\bar{\pi}_{s-\bar{t}}$ with $e_{\bar{t}} = 1$, where $\bar{\pi}_{s'}$ is the probability of being in office s' periods after having won an election, and $\bar{\pi}_0 = 1$. To see this, notice that the voter's belief about α_t will only influence the incumbent's re-election probability in period t, as the incumbent's ability follows an MA(1) process. Thus, only the outcome of an election matters for the incumbent's future survival in office and not the voter's belief prior to the election.

The rest of the proof proceeds in steps. In Step 1, we show that the incumbent will never find it optimal to call for a snap election. Step 2 shows that if the incumbent never calls for a snap election, then it is indeed optimal for incumbent to not call for a snap election when $\alpha_t = \alpha^H$ if and only if the inequality in Proposition 3.1.(2) holds. Given a strategy in which the incumbent calls for a snap election whenever $\alpha_t = \alpha^H$, we re-write $\bar{\pi}_s$ as a function of $\bar{\pi}_{s-1}$ and $\bar{\pi}_{s-2}$. We show this in Step 3 use it to provide a lower bound for the probability of being in office in period $t+s$. In Step 4, we show that we can use the lower bound from Step 3 to provide a sufficient condition such that calling for snap election is optimal when $\alpha_t = \alpha^H$.

Step 1: We show that it is never optimal for the incumbent to call for a snap election when $\alpha_t = \alpha^L$. As there is no re-election concern, the incumbent's policy choice does not depend on whether there is an election. Using $\bar{\pi}_s$, we can re-write the condition for snap

elections to be optimal for the incumbent with $\alpha_t = \alpha^L$ as

$$[\Omega^L - \Omega^O + X](1 - \pi^L) \leq X\left(\pi^L - \beta\pi^\rho\right)\sum_{s=1}\beta^{s+1}\bar{\pi}_s - \beta X\pi^\rho + X\pi^L \lim_{s\to\infty}\beta^{s+1}\bar{\pi}_s,$$

where $\lim_{s\to\infty}\beta^s\bar{\pi}_s = 0$ as $\beta \in (0,1)$ and $\bar{\pi}_{s'} \geq \bar{\pi}_{s'+1}$ for every $s' \in \mathbb{N}$. Because the incumbent has strong office motives, a necessary condition for the incumbent to call for a snap election when $\alpha_t = \alpha^L$ is that the RHS is strictly positive. We re-write the RHS below.

$$\begin{aligned}
&X\left(\pi^L - \beta\pi^\rho\right)\sum_{s=0}\beta^{s+1}\bar{\pi}_s - \beta X\pi^\rho \\
&= -\rho(\pi^H - \pi^L)\left(\beta + \sum_{s=1}^{\infty}\beta^{s+1}\bar{\pi}_s\right) - \beta\pi^L + \pi^L(1-\beta)\sum_{s=1}^{\infty}\beta^s\bar{\pi}_s \\
&\leq -\rho(\pi^H - \pi^L)\sum_{s=0}^{\infty}\beta^{s+1}\bar{\pi}_s - \beta\pi^L + \pi^L(1-\beta)\beta\bar{\pi}_1\sum_{s=0}^{\infty}\beta^s \\
&= -\rho(\pi^H - \pi^L)\sum_{s=0}^{\infty}\beta^{s+1}\bar{\pi}_s - \beta\pi^L(1-\bar{\pi}_1) < 0,
\end{aligned}$$

where the first inequality follows from $\bar{\pi}_1 \geq \bar{\pi}_s$ for all $s \geq 1$. The second inequality follows as $\pi^H > \pi^L$ and $\bar{\pi}_1 \leq 1$. Thus, the incumbent never calls for a snap election when $\alpha_t = \alpha^L$.

Step 2: We show that there exists an equilibrium in which the incumbent *never* calls for a snap election if and only if the inequality in Proposition 3.1.(2) holds. To this end, suppose that there exists an equilibrium in which the incumbent never calls for a snap election. From Step 1, we know that the incumbent has no incentive to deviate when $\alpha_t = \alpha^L$. It is strictly optimal for the incumbent to deviate and call for a snap election if and only if

$$\frac{\Omega^H - \Omega^O}{X}(1 - \pi^H) < \left(\pi^H - \beta\pi^\rho\right)\sum_{s=0}^{\infty}\beta^s\bar{\pi}_s - 1.$$

By assumption the incumbent never calls for a snap election, so $\bar{\pi}_s = \bar{\pi}_{s-1}$ for every odd s (there is never an election in two consecutive periods). For every even $s > 1$, we have $\bar{\pi}_s = \bar{\pi}_{s-2}\pi^\rho$. This follows from the fact that an election takes place every second period regardless of the realisation of the incumbent's ability shock in an equilibrium with no snap election. By combining these two observations we re-write the RHS of the inequality

above, as

$$\begin{aligned}
&\left(\pi^H - \beta\pi^\rho\right)\sum_{s=0}^{\infty}\beta^s\bar{\pi}_s - 1 = \left(\pi^H - \beta\pi^\rho\right)(1+\beta)\sum_{s=0}^{\infty}\beta^{2s}\bar{\pi}_{2s} - 1 \\
&= \left(\pi^H - \beta\pi^\rho\right)(1+\beta)\sum_{s=0}^{\infty}\beta^{2s}\left(\pi^\rho\right)^s - 1 = \frac{\left(\pi^H - \beta\pi^\rho\right)(1+\beta)}{1-\beta^2\pi^\rho} - 1 \\
&= \frac{\pi^H(1+\beta) - (1+\beta\pi^\rho)}{1-\beta^2\pi^\rho}.
\end{aligned}$$

The result follows from noting that $\pi^\rho = \frac{1}{2}$ as $\eta \sim U\left[-\frac{1}{2\psi}, \frac{1}{2\psi}\right]$.

For Step 3 and 4, suppose that there exists an equilibrium in which the incumbent calls for a snap election when $\alpha_t = \alpha^H$, but not when $\alpha_t = \alpha^L$.

Step 3: Given the conjectured strategies we prove by induction that $\bar{\pi}_s$ can be bounded below by $\left(\pi^\rho\right)^{\frac{s}{2}}\left(\rho\pi^H + 1 - \rho\right)^{\frac{s}{2}}$ when s is even and $\left(\pi^\rho\right)^{\frac{s-1}{2}}\left(\rho\pi^H + 1 - \rho\right)^{\frac{s+1}{2}}$ when s is odd.

First, we observe that $\bar{\pi}_s$ can be written as $\bar{\pi}_s = \rho\pi^H\bar{\pi}_{s-1} + (1-\rho)\pi^\rho\bar{\pi}_{s-2}$ for $s \geq 2$. To see this, notice that if the incumbent won an election s periods ago, then she would also face an election $s-1$ periods ago if and only if her utility shock was α^H (which happens with probability ρ). The probability of re-election is then π^H. If she did not have an election $s-1$ periods ago (which happens with probability $1-\rho$), then she would be forced to have an election $s-2$ periods ago (regardless of her ability shock). Her average probability of winning such an election is π^ρ. By combining the above observations we derive the desired expression for $\bar{\pi}_s$.

Using this, we show that we can bound $\bar{\pi}_s$ from below. When $s = 1$, then we know that the incumbent won an election in the previous period: thus if her current period ability shock is α^L, then she will survive with probability one, as there will be no election. If her current period ability shock is α^H, then she will call for a snap election and be re-elected with probability π^H. Thus,

$$\bar{\pi}_1 = \rho\pi^H + 1 - \rho = \left(\pi^\rho\right)^{\frac{s-1}{2}}\left(\rho\pi^H + (1-\rho)\pi^L\right).$$

For $s = 2$, we have

$$\begin{aligned}
\bar{\pi}_2 &= \rho\pi^H\bar{\pi}_1 + (1-\rho)\pi^\rho\bar{\pi}_0 = \pi^\rho\left(\rho\pi^H + 1 - \rho\right) + \rho(1-\rho)\pi^H(1-\pi^L) \\
&\geq \left(\pi^\rho\right)^{\frac{s}{2}}\left(\rho\pi^H + 1 - \rho\right)^{\frac{s}{2}},
\end{aligned}$$

where the inequality follows as $\pi^L < \pi^H \leq 1$.

Inductive step: Assume that the result holds for any $n \leq s-1$. We want to show that is holds for s. If s is even, then

$$\begin{aligned}\bar{\pi}_s &= \rho\pi^H\bar{\pi}_{s-1} + (1-\rho)\pi^\rho\bar{\pi}_{s-2} \\ &\geq \rho\pi^H\left(\pi^\rho\right)^{\frac{s-2}{2}}\left(\rho\pi^H+1-\rho\right)^{\frac{s}{2}} + (1-\rho)\pi^\rho\left(\pi^\rho\right)^{\frac{s-2}{2}}\left(\rho\pi^H+1-\rho\right)^{\frac{s-2}{2}} \\ &= \bar{\pi}_2\left(\pi^\rho\right)^{\frac{s}{2}-1}\left(\rho\pi^H+1-\rho\right)^{\frac{s}{2}-1} \geq \left(\pi^\rho\right)^{\frac{s}{2}}\left(\rho\pi^H+1-\rho\right)^{\frac{s}{2}},\end{aligned}$$

where the inequalities follows from the inductive hypothesis. If s is odd, then

$$\begin{aligned}\bar{\pi}_s &= \rho\pi^H\bar{\pi}_{s-1} + (1-\rho)\pi^\rho\bar{\pi}_{s-2} \\ &\geq \rho\pi^H\left(\pi^\rho\right)^{\frac{s-1}{2}}\left(\rho\pi^H+1-\rho\right)^{\frac{s-1}{2}} + (1-\rho)\pi^\rho\left(\pi^\rho\right)^{\frac{s-3}{2}}\left(\rho\pi^H+1-\rho\right)^{\frac{s-1}{2}} \\ &= \left(\pi^\rho\right)^{\frac{s-1}{2}}\left(\rho\pi^H+1-\rho\right)^{\frac{s+1}{2}},\end{aligned}$$

where the inequality follows from the inductive hypothesis. This completes Step 3.

Step 4: From Step 1, we know that the incumbent has no incentive to deviate from the conjectured strategies when $\alpha_t = \alpha^L$. When $\alpha_t = \alpha^H$ the incumbent will deviate and not call for a snap election if and only if

$$\frac{\Omega^H - \Omega^O}{X}(1-\pi^H) > \left(\pi^H - \beta\pi^\rho\right)\sum_{s=0}\beta^s\bar{\pi}_s - 1 + \pi^H\lim_{s\to\infty}\beta^s\bar{\pi}_s.$$

From Step 3, we have $\bar{\pi}_s + \beta\bar{\pi}_{s+1} \geq \left(\pi^\rho\right)^{\frac{s}{2}}\left(\rho\pi^H+1-\rho\right)^{\frac{s}{2}}\left(1+\beta(\rho\pi^H+1-\rho)\right)$ for every even $s \geq 0$. Using this, and the fact that $\lim_{s\to\infty}\beta^s\bar{\pi}_s = 0$, we can bound the RHS of the inequality as follows

$$\begin{aligned}\left(\pi^H - \beta\pi^\rho\right)\sum_{s=0}^{\infty}\beta^s\bar{\pi}_s - 1 &\geq \left(\pi^H - \beta\pi^\rho\right)\left(1+\beta(\rho\pi^H+1-\rho)\right)\sum_{s=0}^{\infty}\beta^{2s}\left(\pi^\rho\right)^s\left(\rho\pi^H+1-\rho\right)^s - 1 \\ &= \frac{\pi^H\left(1+\beta(\rho\pi^H+1-\rho)\right) - (1+\beta\pi^\rho)}{1-\beta^2\pi^\rho\left(\rho\pi^H+1-\rho\right)}.\end{aligned}$$

This implies that the incumbent has no incentives to deviate if the inequality in Proposition 3.1.(3) holds. This completes the proof.

3.4 Asymmetric Information Case

We now return to the asymmetric information case in which the incumbent has a temporary informational advantage. As the government's investments made in period t only influence the voters' utility in period $t+1$, it is more realistic that the voters cannot observe the period t ability shock before period $t+1$. At which point the voters can infer α_t from k_{t+1} provided they know α_{t-1}. The main result in this Section is Proposition 3.3 which states that snap election can substitute policy distortions, and in some environments this can mitigate the negative effect snap elections have on the voters welfare.

Given an equilibrium, let $(\tau^*(\alpha_{t-1}, \alpha_t, \theta_t), e^*(\alpha_{t-1}, \alpha_t, \theta_t)) = \sigma^I(\alpha_{t-1}, \alpha_t, \theta_t)$ denote the incumbent's equilibrium strategy, and

$$W^*(\alpha_{t-1}, \alpha_t, \theta_t) = U(y - \tau^*(\alpha_{t-1}, \alpha_t, \theta_t)) + \beta V(\alpha_{t-1} + \alpha_t + \tau^*(\alpha_{t-1}, \alpha_t, \theta_t)),$$

be the citizen utility when the incumbent follows the equilibrium strategy in period t given α_{t-1}, α_t and θ_t. We can use the above to express the expected citizen utility from having an incumbent in office in period $t+1$, when the belief that $\alpha_t = \alpha^H$ is $\hat{\rho}$ and $e_t = \theta_{t+1}$:

$$\begin{aligned}\Omega(\hat{\rho}, \theta_{t+1}) = {} & \hat{\rho}\rho W^*(\alpha^H, \alpha^H, \theta_{t+1}) + (1-\hat{\rho})\rho W^*(\alpha^L, \alpha^H, \theta_{t+1}) \\ & + \hat{\rho}(1-\rho) W^*(\alpha^H, \alpha^L, \theta_{t+1}) + (1-\hat{\rho})(1-\rho) W^*(\alpha^L, \alpha^L, \theta_{t+1}).\end{aligned}$$

Given the distribution of η, the probability of being re-elected is

$$\pi_{t,t+1}(\hat{\rho}, 1) = \Pr(\eta_t \geq \Omega(\rho, 1) - \Omega(\hat{\rho}, 1)) = \frac{1}{2} + \psi\left(\Omega(\hat{\rho}, 1) - \Omega(\rho, 1)\right),$$

when the voter's beliefs that $\alpha_t = \alpha^H$ is $\hat{\rho}$. As in the full information case, we denote the probability of election when the voter believes with probability one that $\alpha_t = \alpha^i$ by $\hat{\pi}^i$ for $i = L, H$, and $\hat{\pi}^\rho = \rho\hat{\pi}^H + (1-\rho)\hat{\pi}^L = \frac{1}{2} = \pi^\rho$. Whenever there is no policy distortion in the periods following an election, then $\hat{\pi}^i = \pi^i$ for $i = L, H$.

An incumbent who faces the highest probability of re-election might find it optimal to distort fiscal policies in order to signal her type. For a fixed τ, the difference in the citizen utility for two levels of ability $\varepsilon^H > \varepsilon^L$ is given by $\beta\left(V(\varepsilon^H + \tau) - V(\varepsilon^L + \tau)\right) > 0$, where the inequality follows as V is increasing. Because V is concave, this implies that the marginal cost of decreasing (increasing) taxes is lower for an incumbent with ε^H (ε^L) compared to ε^L (ε^H). Thus, if an incumbent wants to signal a high current ability shock, she does so by distorting taxes downwards. As $\lim_{k\to 0} V(k) = -\infty$, it is always possible

for an incumbent to signal that $\alpha_t = \alpha^H$. We will focus on undominated separating equilibria by imposing the intuitive criterion proposed by Cho and Kreps (1987). The next proposition summarises properties common to every separating equilibrium.

Proposition 3.3. *Given Assumption 3.1. In any separating equilibrium the following properties hold:*

1. $\hat{\pi}^H > \hat{\pi}^L$.
2. *When* $\Omega(\rho, 1) - \Omega(0, 1) < X$, *the constraint facing the incumbent with* $\alpha_t = \alpha^H$ *is strictly relaxed when* $\theta_t = 1$ *compared to when* $\theta_t = 0$.
3. *In this case, fiscal policy and snap elections are signalling substitutes.*
4. *If the incumbent with* $\alpha_t = \alpha^H$ *distorts the fiscal policy to deter the incumbent with* $\alpha_t = \alpha^L$ *from mimicking her for any* $\alpha_{t-1} \in \{\alpha^L, \alpha^H\}$, *the utility loss from fiscal policy distortion is larger when* $\alpha_{t-1} = \alpha^H$ *compared to* $\alpha_{t-1} = \alpha^L$.

Furthermore, there exists no separating equilibrium in which an incumbent with $\alpha_t = \alpha^L$ *calls for a snap election.*

We first provide an outline of the proof, and below we discuss the proposition's implications. The formal proof is delegated to the appendix. If there is no election, then as the incumbent's ability follows an MA(1) process there are no re-election concerns, and thus no incentives for the incumbent to distort her policy choices, i.e. when $e^*(\alpha_{t-1}, \alpha_t, 1) = 0$, then $\tau^*(\alpha_{t-1}, \alpha_t, 1) = \tau^{FI}(\alpha_{t-1} + \alpha_t)$ for any $(\alpha_{t-1}, \alpha_t) \in \{\alpha^L, \alpha^H\}^2$. This implies that in any separating equilibria with no snap election we have $\Omega(\mathbb{1}_{i=H}, 1) = \Omega^i$ for $i = L, H$, and $\hat{\pi}^i = \pi^i$ for $i = L, H$. That is, the probability of re-election is the same as in the full information case. The incumbent may still have an incentive to distort her policy choices in order to influence the voter's inference about her ability shock in election periods. Thus, in a separating equilibrium with snap election the incumbent may distort her policy choices in the periods immediately following an election. In which case, it may influence the probability of re-election, i.e. $\hat{\pi}^i \neq \pi^i$ for $i = L, H$. We therefore need to show that in every separating equilibrium $\hat{\pi}^H > \hat{\pi}^L$. In order to do this, we first show that if $\Omega(\rho, 1) - \min\{\Omega(0, 1), \Omega(1, 1)\} < X$ and $\Omega(1, 1) > (<)\Omega(0, 1)$ then the incumbent with $\alpha_t = \alpha^L$ ($\alpha_t = \alpha^H$) has a strictly lower incentive to mimic when $\theta_t = 1$ compared to when $\theta_t = 0$. This follows from similar arguments as in Step 1 of Proposition 3.1. Second, if $\Omega(\rho, 1) - \min\{\Omega(0, 1), \Omega(1, 1)\} > X$ and $\Omega(1, 1) > (<)\Omega(0, 1)$, then the utility loss from distortion is largest when $\alpha_{t-1} = \alpha^L$ ($\alpha_{t-1} = \alpha^H$). Third, we show that $\hat{\pi}^H > \hat{\pi}^L$ in every

separating equilibrium in which the incumbent with the lowest probability of re-election does not call for a snap election. Afterwards, we show that there exists no separating equilibrium in which the incumbent who faces the lowest probability of re-election calls for a snap election. Thus, establishing that $\hat{\pi}^H > \hat{\pi}^L$ in every separating equilibrium. Finally, we show that when the constraint is binding for an incumbent with $\alpha_t = \alpha^H$ regardless of the realisation of α_{t-1}, then the utility loss from distortion is largest for the incumbent with $\alpha_{t-1} = \alpha^H$.

Proposition 3.3.(1) implies that the incumbent will only distort her policy choices when $\alpha_t = \alpha^H$. The incumbent with $\alpha_t = \alpha^L$ has higher incentives to mimic an incumbent with $\alpha_t = \alpha^H$ when she is forced to have an election. Thus, under a flexible election regime the incumbent has two signalling tools, and fiscal policy distortions are less pronounced in periods with snap elections. This means that the detrimental impacts of a flexible election regime we observed in the full information case is somewhat mitigated in the asymmetric information case. Furthermore, when the incumbent with $\alpha_t = \alpha^H$ faces a binding constraint regardless of whether $\alpha_{t-1} = \alpha^L$ or α^H, then the utility loss from policy distortion is larger when $\alpha_{t-1} = \alpha^H$. When this happens in periods with a regular election (and there is no distortion in periods with a snap election), there is a further mitigation of the negative consequences of a flexible election regime. In particular, the flexible election regime improves the citizens' utility, if such an equilibrium exists and the following expression is strictly positive:

$$
\begin{aligned}
&\frac{\rho^2(1-\rho)}{2(2-\rho)}\left(W^{FI}(2\alpha^H) - W^*(\alpha^H,\alpha^H,0) - \left(W^{FI}(\alpha^L+\alpha^H) - W^*(\alpha^L,\alpha^H,0)\right)\right)\\
&+\frac{\rho^2}{2(2-\rho)}\left(W^{FI}(2\alpha^H) - W^*(\alpha^H,\alpha^H,0)\right)\\
&+\frac{\rho^2\left(\frac{1}{2}\rho-(3\rho-2)(1-\pi^H)\right)}{2(2-\rho)}\left(W^{FI}(\alpha^H+\alpha^L) - W^{FI}(2\alpha^L) - \left(W^{FI}(2\alpha^H) - W^{FI}(\alpha^H+\alpha^L)\right)\right)\\
&-\frac{\rho\left(\frac{1}{2}\rho-(3\rho-2)(1-\pi^H)\right)}{2(2-\rho)}\left(W^{FI}(\alpha^H+\alpha^L) - W^{FI}(2\alpha^L)\right),
\end{aligned}
$$

where the weights are derived using the distributions of the incumbent's ability from the end of Section 3.3. The first three terms are positive and the last term is negative: The first term is positive by Proposition 3.3.(4), the second term is trivially positive as policy distortion leads to a utility loss and the third (fourth) term is positive (negative) as W^{FI} is an increasing concave function and $\frac{1}{2}\rho-(3\rho-2)(1-\pi^H) > 0$ for any $\pi^H \in \left(\frac{1}{2},1\right]$ and $\rho \in (0,1)$. Whether there exists an equilibrium in which flexible election timing improves welfare remains an open question.

In equilibria in which fiscal policies are distorted in periods with snap election the comparison to the fixed election regime becomes more complicated, as the tax level influences the incumbent's re-election probabilities above and beyond what it signals about the incumbent's ability. In particular, if the constraints in snap elections are binding when $(\alpha_{t-1}, \alpha_t) = (\alpha^H, \alpha^H)$ and $(\alpha_{t-1}, \alpha_t) = (\alpha^L, \alpha^H)$, then the difference in re-election probabilities decreases compared to the full information case. A decrease in $\hat{\pi}^H - \hat{\pi}^L$ decreases the incumbent's incentive to mimic an incumbent with $\alpha_t = \alpha^H$, but further strengthen the perverse effects snap elections has on the long run distribution of the incumbent's ability.

In addition the parameter values for which separating equilibria with and without snap elections exist differs compared to the full information benchmark. The next proposition provides sufficient conditions for the existence of a separating equilibrium in which the incumbent never calls for a snap election and when the incumbent with $\alpha_t = \alpha^H$ always calls for a snap election.

Proposition 3.4. *Given Assumption 3.1, there exists a separating equilibrium in which*

1. *the incumbent never calls for a snap election if*

$$\frac{\Omega^H - \Omega^O}{X}(1 - \pi^H) - \frac{\Omega^H - \Omega(1,0)}{X} \geq \frac{2(1+\beta)\pi^H - (2+\beta)}{2 - \beta^2}$$

2. *the incumbent with $\alpha_t = \alpha^H$ calls for a snap election if*

$$\frac{W^{FI}(2\alpha^H) - W^*(\alpha^H, \alpha^H, 1)}{\beta X} + \frac{\Omega(1,1) - \Omega(\rho,1)}{X}(1 - \hat{\pi}^H) - \frac{\Omega(1,1) - \Omega(1,0)}{X}$$
$$\leq \frac{2\hat{\pi}^H\left(1 + \beta[\rho\hat{\pi}^H + 1 - \rho]\right) - (2+\beta)}{2 - \beta^2\left(1 + \beta[\rho\hat{\pi}^H + 1 - \rho]\right)}$$

and when $\theta_t = 1$ the utility loss from policy distortion is weakly larger when $\alpha_{t-1} = \alpha^H$.

In an equilibrium without snap elections, the benefits of calling a snap election depends on the voter's off-equilibrium beliefs. An incumbent with $\alpha_t = \alpha^H$ will achieve the highest utility from deviating and calling a snap election, if this leads the voter to believe that $\alpha_t = \alpha^H$ without any fiscal policy distortion. In this case the incumbent with $\alpha_t = \alpha^H$ prefers to not call for a snap election if and only if the inequality in Proposition 3.4.(1) holds. Notice that the only difference between this inequality and the one in Proposition 3.1.(2) is the additional term on the RHS. This term accounts for the fact that the citizen

utility is higher in the periods just after an election for a given level of ability, as there is less utility loss due to policy distortions. The inequality in Proposition 3.4.(2) differs from the inequality in three ways. The first term on the LHS comes from the immediate gain of no policy distortion when incumbent does not call for a snap election. As above, the third term on the LHS stems from the fact that the policy distortion is lower in periods with a snap election relative to periods in which elections have to take place. Finally, as the utility loss from policy distortions is largest when $\alpha_t = \alpha^H$, this implies that in face of policy distortions prior to snap elections we have $\Omega(1,1) - \Omega(\rho,1) < \Omega^H - \Omega^O$ and consequently $\hat{\pi}^H < \pi^H$.

3.5 Conclusion

This paper analyses the impact of having a flexible election regime. I show that allowing the incumbent to opportunistically call for snap elections has detrimental effect on welfare in the benchmark with full information. When the incumbent has a temporary informational advantage compared to the voters, the flexible election timing provides the incumbent with two signalling tools. Whenever the ego-rents are high enough such that even an incumbent who is relatively unsuited to deal with next year's challenges still prefers to be in office the two tools are signalling are substitutes. This mitigate some of the detrimental effects from having a flexible election regime in the full information case. Whether a flexible election regime ever improve welfare remains an open question.

Even in the bench mark case of full information snap election changes the political budget cycles in equilibrium. If there exists a separating equilibrium in which any incumbent with calls for a snap elections when her current ability shock is high and her fiscal policies are distorted, then the political budget cycles are less pronounced under a flexible election regime.

References

Lasse Aaskoven. The electoral cycle in political contributions: The incumbency advantage of early elections. *Acta Politica*, 55:670–691, July 2020. doi: https://doi.org/10.1057/s41269-019-00138-3.

Alberto Alesina and Roberto Perotti. Fiscal adjustments in OECD countries: Composition and macroeconomic effects. *International Monetary Funds*, 44(2):210–248, June 1997. URL `https://www.jstor.org/stable/3867543`.

Alberto Alesina and Nouriel Roubini. Political cycles in OECD economies. *The Review of Economic Studies*, 59:663–688, October 1992.

Alberto Alesina, Gerald D. Cohen, and Nouriel Roubini. Electoral business cycle in industrial democracies. *European Journal of Political Economy*, 9(1):1–23, March 1993. doi: https://doi.org/10.1016/0176-2680(93)90027-R.

Rui Nuno Baleiras and Vasco Santos. Behavioral and institutional determinants of political business cycles. *Public Choice*, 104(1):121–147, July 2000. URL `https://www.jstor.org/stable/30026465`.

Nathan S Balke. The rational timing of parliamentary elections. *Public Choice*, 65:201–216, 1990.

David P Baron. Comparative dynamics of parliamentaty governments. *American Political Science Review*, 92(3):593–609, September 1998. doi: https://doi.org/10.2307/2585483.

Andre Blais and Richard Nadeau. The electoral budget cycle. *Public Choice*, 74(4):389–403, 1992. URL `https://www.jstor.org/stable/30025623`.

Adi Brender and Allan Drazen. Political budget cycles in new versus established democracies. *Journal of Monetary Economics*, 52:1271–1295, October 2005. doi: 10.1016/j.jmoneco.2005.04.004.

Brandice Canes-Wrone and Jee-Kwang Park. Electoral business cycles in OECD countries. *American Political Science Review*, 106(1):103–122, 2012.

Thomas F. Cargill and Michael M. Hutchinson. Political business cycles with endogenous election timing: Evidence from Japan. *The Review of Economics and Statistics*, 73(4): 733–739, November 1991. URL https://www.jstor.org/stable/2109416.

Fredrik Carlsen. Opinion polls and political business cycles: Theory and evidence for the United States. *Public Choices*, 92:387–406, 1997.

Fredrik Carlsen. Inflation and elections: Theory and evidence for six OECD economies. *Economic Inquiry*, 37(1):120–135, July 2007. doi: https://doi.org/10.1111/j.1465-7295.1999.tb01420.x.

D. Chappell and D. A. Peel. On the political theory of the business cycle. *Economics Letters*, 2:327–332, 1979.

Abdur R. Chowdhury. Political surfing over economic waves: Parliamentary election timing in India. *American Journal of Political Science*, 37(4):1100–1118, November 1993. URL https://www.jstor.org/stable/2111545.

Maria de los Angeles Gonzalez. Do changes in democrazy affect the political budget cycle? evidence from Mexico. *Review of Development Economics*, 6(2):204–224, 2002.

Allan Drazen. The political business cycle after 25 years. *NBER Macroeconomics Annual*, 15:75–117, 2000. doi: https://doi.org/10.1086/654407.

Eric Dubois. Political business cycles 40 years after Nordhaus. *Public Choice*, 166:235–259, February 2016. doi: 10.1007/s11127-016-0313-z.

J. Stephen Ferris and Derek E. H. Olmstead. Fixed versus flexible election terms: explaining innovation in the timing of Canada's election cycle. *Constitutional Political Economy*, 28:117–141, February 2017. doi: 10.1007/s10602-017-9237-y.

Victor Ginsburgh and Phillippe Michel. Random timing of elections and the political business cycle. *Public Choice*, 40(2):155–164, 1983. URL https://www.jstor.org/stable/30023659.

Kevin B. Grier. On the existence of a political monetary cycle. *American Journal of Political Science*, 33(2):376–389, May 1989. doi: https://doi.org/10.2307/2111152.

Sanford Grossman and Oliver Hart. An analysis of the principal-agent problem. *Econometrica*, 51:7–45, 1983.

Jac C. Heckelman and Hakan Berument. Political business cycles and endogenous elections. *Southern Economic Journal*, 64(4):987–1000, April 1998. URL https://www.jstor.org/stable/1061215.

Takatoshi Ito. The timing of elections and political business cycles in Japan. *Journal of Asian Economies*, 1(1):135–156, 1990.

Takatoshi Ito and Jin Hyuk Park. Political business cycles in the parlimentary system. *Economics Letters*, 27:233–238, 1988. doi: https://doi.org/10.1016/0165-1765(88)90176-0.

Mark Andreas Kayser. Who surfs, who manipulates? the determinants of opportunistic election timing and electorally motivated economic intervention. *The American Political Science Review*, 99(1):17–27, February 2005. URL `https://www.jstor.org/stable/30038916`.

Ulrich Lachler. On political business cycles with endogenous election dates. *Journal of Public Economics*, 17:111–117, 1982. doi: https://doi.org/10.1016/0047-2727(82)90029-9.

Susanne Lohnmann. Rationalizing the political business cycles: A workhorse model. *Economics and Politics*, 10(1):1–17, March 2003. doi: https://doi.org/10.1111/1468-0343.00035.

Arthur Lupia and Kaare Strom. Coalition termination and the strategic timing of parliamentary elections. *The American Political Science Reivew*, 89(3):648–665, September 1995. URL `https://www.jstor.org/stable/2082980`.

William Nordhaus. The political budget cycles. *Review of Economic Studies*, 42(2):169–190, April 1975.

Harvey D Palmer and Guy D. Whitten. Government compentence, economic performance and endogenous election dates. *Electoral Studies*, 19:413–426, June 2000. doi: https://doi.org/10.1016/S0261-3794(99)00059-1.

Bradford G. Reid. Endogenous elections, electoral budget cycles and Canadian provincial governments. *Public Choice*, 97:35–48, October 1998.

Kenneth Rogoff. Equilibrium political budget cycles. *The American Economic Review*, 80 (1):21–36, March 1990. URL `https://www.jstor.org/stable/2006731`.

Kenneth Rogoff and Anne Sibert. Elections and macroeconomic policy cycles. *The Review of Economic Studies*, 55(1):1–16, January 1988. doi: https://doi.org/10.2307/2297526.

Petra Schleiter and Margit Tavits. The electoral benefits of opportunistic election timing. *The Journal of Politics*, 78(3):837–850, July 2016. doi: http://dx.doi.org/10.1086/685447.

Christina J. Schneider. Fighting with one hand tied behind the back: Political budget cycles in the West German States. *Public Choice*, 142(1):125–150, January 2010. URL http://www.jstor.com/stable/40541952.

Kenneth A. Schultz. The politics of the political business cycle. *British Journal of Political Science*, 25(1):79–99, January 1995. URL https://www.jstor.org/stable/194177.

Alastair Smith. Election timing in majoritarian parliaments. *British Journal of Political Science*, 33(3):397–418, July 2003. URL https://www.jstor.org/stable/4092304.

Alastair Smith. *Election Timing*. Cambridge University Press, 2004.

Laron K. Williams. Flexible election timing and international conflict. *International Studies Quarterly*, 57(3):449–461, September 2013. URL https://www.jstor.org/stable/24017916.

3.6 Appendix

Proof of Lemma 3.1. Consider an equilibrium in which the incumbent calls for a snap election if and only if $\alpha_t = \alpha^H$. We observe that given $\nu_{t-1}, \delta_{t-1}, \xi^i_{t-1}$ for $i = L, H$ the incumbent's ability shock (α_{t-1}, α_t) is given by

- (α^H, α^H) with probability

  $\left(\nu_{t-1} + \xi^H_{t-1}\right)\pi^H\rho + \left(\left(\nu_{t-1} + \xi^H_{t-1}\right)(1-\pi^H) + \xi^L_{t-1}(1-\pi^L)\right)\rho^2,$

- (α^H, α^L) with probability

  $\left(\nu_{t-1} + \xi^H_{t-1}\right)\pi^H(1-\rho) + \left(\left(\nu_{t-1} + \xi^H_{t-1}\right)(1-\pi^H) + \xi^L_{t-1}(1-\pi^L)\right)\rho(1-\rho),$

- (α^L, α^H) with probability

  $\delta_{t-1}\rho + \xi^L_{t-1}\pi^L\rho + \left(\left(\nu_{t-1} + \xi^H_{t-1}\right)(1-\pi^H) + \xi^L_{t-1}(1-\pi^L)\right)(1-\rho)\rho,$ and

- (α^L, α^L) with probability

  $\delta_{t-1}(1-\rho) + \xi^L_{t-1}\pi^L(1-\rho) + \left(\left(\nu_{t-1} + \xi^H_{t-1}\right)(1-\pi^H) + \xi^L_{t-1}(1-\pi^L)\right)(1-\rho)^2.$

To see this, notice that if the incumbent has $\alpha_{t-1} = \alpha^H$, then she will face an election in period $t-1$. Thus, she will become a (α^H, α^H) incumbent if she wins the election and nature draws $\alpha_t = \alpha^H$: $\left(\nu_{t-1} + \xi^H_{t-1}\right)\pi^H\rho$. If the opponent comes into office after an election in period $t-1$, then the incumbent in period t will have (α^H, α^H) with probability ρ^2. The probability that an opponent will take over the office in period t is $\left(\left(\nu_{t-1} + \xi^H_{t-1}\right)(1-\pi^H) + \xi^L_{t-1}(1-\pi^L)\right)$. The probability that the incumbent in period t has (α^H, α^L) follows the same reasoning. When the incumbent has $\alpha_{t-1} = \alpha^L$, there is an additional term as she will not call for a snap election, and thus is sure to continue in office in the next period.

Given the equilibrium, the only reason for a snap election is if there was an election in the previous period and $\alpha_t = \alpha^H$. Thus, we have $\nu_t = (1-\delta_{t-1})\rho$. If there was no election in the previous period, the incumbent will be forced to call an election in the current period regardless of the realisation of α_t. As α_t is independent of whether there was an election in the previous period, we have $\xi^H_t = \delta_{t-1}\rho$ and $\xi^L_t = \delta_{t-1}(1-\rho)$. Finally, if there was an election in the previous period and $\alpha_t = \alpha^L$, then there will be no election in period t: $\delta_t = (1-\delta_{t-1})(1-\rho)$.

Observation 3.1. *$\delta_t = (1-\rho)\sum_{s=0}^{t-1}(\rho-1)^s$ for every $t \in N$.*

Proof. We prove this by induction. Since there is no forced election in period 1, this trivially holds for $t = 1$. Suppose that it holds for every $t' < t$. We have

$$\begin{aligned}\delta_t = (1-\delta_{t-1})(1-\rho) &= \left(1-(1-\rho)\sum_{s=0}^{t-1}(\rho-1)^s\right)(1-\rho) \\ &= \left(1-\sum_{s=1}^{t}(\rho-1)^s\right)(1-\rho) = (1-\rho)\sum_{s=0}^{t-1}(\rho-1)^s\end{aligned}$$

□

By observation 3.1 we have $\tilde{\delta} = \lim_{t\to\infty}(1-\rho)\sum_{s=0}^{t-1}(\rho-1)^s = (1-\rho)\sum_{s=0}^{\infty}(\rho-1)^s = \frac{1-\rho}{2-\rho}$, where the last equallity follows since $|\rho-1| < 1$. Using the above, we derive $\tilde{\nu} = \frac{\rho}{2-\rho}$, $\tilde{\xi}^H = \frac{\rho(1-\rho)}{2-\rho}$ and $\tilde{\xi}^L = \frac{(1-\rho)^2}{2-\rho}$. This completes the proof. □

Proof of Proposition 3.2. Under the fixed election regime the distribution of the incumbent's ability does not converge, but alternates between the distribution immediately after an election and the distribution prior to any election. In the long run the expected citizen utility over a business cycle is given by:

$$E^{fixed} = \delta_1 W^{FI}(2\alpha^H) + \delta_2 W^{FI}(\alpha^H + \alpha^L) + \delta_3 W^{FI}(2\alpha^L),$$

where $\delta_1 = \rho^2\left[\pi^H + \frac{1}{2} + \beta\right]$, $\delta_2 = \rho(1-\rho)\left[1 + \pi^H + \pi^L + 2\beta\right]$ and $\delta_3 = (1-\rho)^2\left[\pi^L + \frac{1}{2} + \beta\right]$. The long run the expected citizen utility (averaging over the two periods) is given by:

$$E^{flex} = \gamma_1 W^{FI}(2\alpha^H) + \gamma_2 W^{FI}(\alpha^H + \alpha^L) + \gamma_3 W^{FI}(2\alpha^L),$$

where $\gamma_1 = \frac{\rho^2(1+\beta)}{2(2-\rho)}\left[1 + \rho 4\pi^H(1-\rho)\right]$, $\gamma_2 = \frac{\rho(1-\rho)(1+\beta)}{2(2-\rho)}\left[5 + 2\rho + 4\pi^H - 8\rho\pi^H\right]$ and $\gamma_3 = \frac{(1-\rho)^2(1+\beta)}{2(2-\rho)}\left[4 - 4\rho\pi^H + \rho\right]$. We can write the difference between the citizen utility under the fixed and flexible election regime as:

$$\begin{aligned}&E^{fixed} - E^{flex} = \\ &(\delta_1 - \gamma_1)\left[W^{FI}(2\alpha^H) - W^{FI}(\alpha^H + \alpha^L)\right] + (\gamma_3 - \delta_3)\left[W^{FI}(\alpha^H + \alpha^L) - W^{FI}(2\alpha^L)\right],\end{aligned}$$

as $\delta_1 + \delta_2 + \delta_3 = 1$ and $\gamma_1 + \gamma_2 + \gamma_3 = 1$. Thus, if $\delta_1 - \gamma_1 > 0$ and $\gamma_3 - \delta_3 > 0$, this completes the proof.

First consider $\delta_1 - \gamma_1$:

$$\begin{aligned}\delta_1 - \gamma_1 &= \frac{\rho^2}{2(2-\rho)}\left[4\pi^H - 2\rho\pi^H + 2 - \rho + 4\beta - 2\beta\rho - [1+\beta](1+\rho+4\pi^H(1-\rho))\right] \\ &= \frac{\rho^2}{2(2-\rho)}\left[1 - 2\rho(1-\pi^H) + \beta(1-\rho)[3-4\pi^H]\right].\end{aligned}$$

As $\rho \in (0,1)$ and $\pi^H > \frac{1}{2}$, then $1 - 2\rho(1-\pi^H) > 0$. Thus, if $\pi^H \leq \frac{3}{4}$ then $\delta_1 - \gamma_1 > 0$. Therefore suppose this is not the case. We can bound $\delta_1 - \gamma_1$ from below by

$$\delta_1 - \gamma_1 > \frac{\rho^2}{2(2-\rho)}\left[1 - 2\rho(1-\pi^H) + (1-\rho)[3-4\pi^H]\right].$$

If $\pi^H = \frac{1}{2}$, then $\delta_1 - \gamma_1 > \frac{2\rho^2(1-\rho)}{2(2-\rho)} > 0$. If $\pi^H = 1$, then $\delta_1 - \gamma_1 > \frac{\rho^3}{2(2-\rho)} > 0$. As $\delta_1 - \gamma_1$ is linear in π^H and $\pi^H \in \left(\frac{1}{2}, 1\right]$, we conclude that $\delta_1 - \gamma_1 > 0$. Next, we show that $\gamma_3 - \delta_3 > 0$.

$$\begin{aligned}\gamma_3 - \delta_3 &= \frac{(1-\rho)^2}{2(2-\rho)}\left[[1+\beta](4-4\rho\pi^H+\rho) - (2(2-\rho)\pi^L + (2-\rho) + 2\beta(2-\rho))\right] \\ &= \frac{(1-\rho)}{2(2-\rho)}\left[\rho(1-2\rho(1-\pi^H)) + \beta\rho(1-\rho)(3-4\pi^H)\right],\end{aligned}$$

where the second equality follows as $\rho\pi^H + (1-\rho)\pi^L = \frac{1}{2}$. If $\pi^H = \frac{1}{2}$, then $\gamma_3 - \delta_3 = \frac{[1+\beta]\rho(1-\rho)^2}{2(2-\rho)} > 0$, and if $\pi^H = 1$, then $\gamma_3 - \delta_3 = \frac{\rho(1-\rho)}{2(2-\rho)}[1-\beta(1-\rho)] > 0$. This completes the proof. □

3.6.1 Proof of Proposition 3.3

First, we observe that when there is no election, i.e. $e_t = 0$, then there is no re-election concerns and thus no policy distortion. So $\tau^*(\alpha_{t-1}, \alpha_t, \theta_t) = \tau^{FI}(\alpha_{t-1}, \alpha_t)$ whenever $e^*(\alpha_{t-1}, \alpha_t, \theta_t) = 0$. This implies that in any separating equilibria in which there are no snap elections, $\Omega(1,1) = \Omega^H > \Omega^L = \Omega(0,1)$ and $\hat{\pi}^H = \pi^H > \pi^L = \hat{\pi}^L$.

Similarly, the incumbent facing the lowest probability of re-election has no incentives the distort fiscal policy, as there is no risk of being mimicked. That is, $\tau^*(\alpha_{t-1}, \alpha^i, \theta_t) = \tau^{FI}(\alpha_{t+1}, \alpha^i)$ whenever $\Omega(\mathbb{1}_{i=H}, 1) \leq \Omega(\mathbb{1}_{i\neq H}, 1)$.

The remainder of the proof proceeds in steps. In Step 1, we show that if $\Omega(\rho,1) - \min\{\Omega(0,1), \Omega(1,1)\} < X$ and $\Omega(1,1) > (<)\Omega(0,1)$, then the constraint facing the incumbent with $\alpha_t = \alpha^H$ ($\alpha_t = \alpha^L$) is strictly relaxed when $\theta_t = 1$ compared to when $\theta_t = 0$. Step 2 shows that if $\Omega(\rho,1) - \min\{\Omega(0,1), \Omega(1,1)\} \geq X$ and $\Omega(1,1) > (<)\Omega(0,1)$, then the utility loss from distortion is largest when $\alpha_{t-1} = \alpha^L$ ($\alpha_{t-1} = \alpha^H$). Step 3 shows that

in every separating equilibrium in which the incumbent who faces the lowest probability of re-election does not call for a snap election we have $\Omega(1,1) > \Omega(0,1)$. Step 4 asserts that there exists no separating equilibrium in which the incumbent who faces the lowest probability of re-election calls for a snap election. By combining Step 3 and 4 it implies that $\Omega(1,1) > \Omega(0,1)$ and hence $\hat{\pi}^H > \hat{\pi}^L$ in every separating equilibrium. Furthermore, there exist no separating equilibria in which an incumbent with $\alpha_t = \alpha^L$ calls for a snap election. Finally, Step 5 shows that in any election in which the constraints are binding for $(\alpha_{t-1}, \alpha_t) = (\alpha^H, \alpha^H)$ and $(\alpha_{t-1}, \alpha_t) = (\alpha^L, \alpha^H)$ the utility loss from distortion is highest when $\alpha_{t-1} = \alpha^H$, which concludes the proof.

Step 1: Suppose that $\Omega(\rho,1) - \min\{\Omega(0,1), \Omega(1,1)\} < X$ and $\Omega(1,1) > (<)\Omega(0,1)$. We show that the constraint facing the incumbent with the highest probability of re-election is relaxed when $\theta_t = 1$ compared to when $\theta_t = 0$.

Let $\Omega(\mathbb{1}_{i=H}, 1) > \Omega(\mathbb{1}_{j=H}, 1)$ for $i \in \{L, H\}$ and $j \neq i$. The constraint facing the incumbent with $\alpha_t = \alpha^i$ is not relaxed when $\theta_t = 1$ compared to when $\theta_t = 0$ if

$$\Omega(\mathbb{1}_{j=H},1)\hat{\pi}^j + \Omega(\rho,1)(1-\hat{\pi}^j) + \hat{\pi}^j X \sum_{s=0}^{\infty} \beta^s \bar{\pi}_s \geq \Omega(\mathbb{1}_{j=H},0) + X + \beta\hat{\pi}^\rho X \sum_{s=0}^{\infty} \beta^s \bar{\pi}_s,$$

or equivalently

$$\begin{aligned}
&\left(\Omega(\mathbb{1}_{j=H},1) - \Omega(\rho,1) + X\right)(1-\hat{\pi}^j) \\
&\leq X\left(\hat{\pi}^j - \beta\hat{\pi}^\rho\right)\sum_{s=1}^{\infty}\beta^s\bar{\pi}_s - \beta X\hat{\pi}^\rho + \left(\Omega(\mathbb{1}_{j=H},1) - \Omega(\mathbb{1}_{j=H},0)\right).
\end{aligned}$$

By assumption the LHS of the inequality is positive. By the same arguments as in Step 1 of Proposition 3.1 $X\left(\hat{\pi}^j - \beta\hat{\pi}^\rho\right)\sum_{s=1}^{\infty}\beta^s\bar{\pi}_s - \beta X\hat{\pi}^\rho < 0$. Thus, a necessary condition for the constraint not to be relaxed when $\theta_t = 1$ is $\Omega(\mathbb{1}_{i=H},1) > \Omega(\mathbb{1}_{j=H},0)$. This implies that the there is more distortion when $\theta_t = 0$ than when $\theta_t = 1$. However, this only happens if the constraint facing the incumbent with the highest probability of re-election is relaxed when $\theta_t = 1$ compared to $\theta_t = 0$. This concludes the proof of Step 1.

Step 2: We show that if $\Omega(\rho,1) - \min\{\Omega(0,1), \Omega(1,1)\} > X$ and $\Omega(1,1) > (<)\Omega(0,1)$, then the utility loss from distortion is largest when $\alpha_{t-1} = \alpha^L$ ($\alpha_{t-1} = \alpha^H$).

First, assume that there exists a separating equilibrium in which $\Omega(1,1) > \Omega(0,1)$ and $\Omega(\rho,1) - \Omega(0,1) \geq X$. This implies that

$$\rho\left(\Omega(1,1) - \Omega(0,1)\right) = \Omega(\rho,1) - \Omega(0,1) \geq X > \Omega^O - \Omega^L = \rho\left(\Omega^H - \Omega^L\right).$$

As $\Omega(1,1) > \Omega(0,1)$, an incumbent with $\alpha_t = \alpha^L$ does not distort her fiscal policy choices. Therefore, we can re-write the above as

$$W^{FI}(\alpha^L + \alpha^H) - W^*(\alpha^L, \alpha^H, 1) > W^{FI}(2\alpha^H) - W^*(\alpha^H, \alpha^H, 1).$$

This proves the claim.

Assume, now that there exists a separating equilibrium $\Omega(0,1) > \Omega(1,1)$ and $\Omega(\rho,1) - \Omega(1,1) \geq X$. This implies that

$$(1-\rho)\left(\Omega(0,1) - \Omega(1,1)\right) = \Omega(\rho,1) - \Omega(1,1) \geq X > \Omega^O - \Omega^L = \rho\left(\Omega^H - \Omega^L\right).$$

As $\Omega(1,1) < \Omega(0,1)$, an incumbent with $\alpha_t = \alpha^H$ does not distort her fiscal policy choices. Therefore, we can re-write the above as

$$(1-\rho)\left(W^{FI}(\alpha^H + \alpha^L) - W^*(\alpha^H, \alpha^L, 1) - \left(W^{FI}(2\alpha^L) - W^*(\alpha^L, \alpha^L, 1)\right)\right) > \Omega^H - \Omega^L > 0,$$

which implies that the utility loss is largest when $\alpha_{t-1} = \alpha^H$.

This concludes the proof of Proposition 3.3.(2) and 3.3.(3).

Step 3: Consider a separating equilibrium in which the incumbent who faces the lowest probability of re-election does not call for a snap election. Assume for contraction that $\Omega(1,1) \leq \Omega(0,1)$. As the incumbent with $\alpha_t = \alpha^H$ faces the lowest probability of re-election, she has no incentive to distort her policy choices. Thus, $\Omega(1,1) \leq \Omega(0,1)$ implies that

$$W^*(\alpha^H, \alpha^L, 1) < W^*(\alpha^L, \alpha^L, 1) \tag{3.4}$$

As $W^{FI}(\alpha^H + \alpha^L) > W^{FI}(2\alpha^L)$, the constraint is binding for an incumbent with $(\alpha_{t-1}, \alpha_t) = (\alpha^H, \alpha^L)$ when $\theta_t = 1$, and $\tau^*(\alpha^H, \alpha^L, 1) > \tau^{FI}(\alpha^L + \alpha^H)$. Therefore,

$$\begin{aligned} &W^{FI}(2\alpha^H) - U(y - \tau^*(\alpha^L, \alpha^H, 1)) - \beta V(2\alpha^H + \tau^*(\alpha^L, \alpha^H, 1)) \qquad (3.5)\\ &= A \leq W^{FI}(\alpha^L + \alpha^H) - U(y - \tau^*(\alpha^L, \alpha^L, 1)) - \beta V(\alpha^L + \alpha^H + \tau^*(\alpha^L, \alpha^L, 1)), \end{aligned}$$

where $A = \beta\left(\hat{\pi}^L - \hat{\pi}^H\right)\left(\Omega(1,1) - \Omega(\rho,1) + X\right) + \beta\left(\hat{\pi}^L - \hat{\pi}^H\right) X \sum_{s=1}^{\infty} \beta^s \bar{\pi}_s$ is the expected gain from mimicking the incumbent with $\alpha_t = \alpha^L$. Inequality (3.5) can be re-

arranged to

$$\begin{aligned} W^*(\alpha^L, \alpha^L, 1) - W^*(\alpha^L, \alpha^H, 1) \leq W^{FI}(\alpha^H + \alpha^L) - W^{FI}(2\alpha^H) \\ + \beta \left(V(2\alpha^H + \tau^*(\alpha^H, \alpha^L, 1)) - V(\alpha^H + \alpha^L + \tau^*(\alpha^H, \alpha^L, 1)) \right) \\ - \beta \left(V(\alpha^H + \alpha^L + \tau^*(\alpha^L, \alpha^L, 1)) - V(2\alpha^L + \tau^*(\alpha^L, \alpha^L, 1)) \right). \end{aligned}$$

By inequality (3.4), the LHS is strictly positive, and $W^{FI}(\alpha^H + \alpha^L) - W^{FI}(2\alpha^H) < 0$ as W^{FI} is increasing. As V is concave, we have an immediate contradiction unless

$$\alpha^H + \tau^*(\alpha^H, \alpha^L, 1) < \alpha^L + \tau^*(\alpha^L, \alpha^L, 1) \tag{3.6}$$

So suppose that inequality (3.6) holds. This implies that

$$0 < \alpha^H - \alpha^L < \tau^*(\alpha^L, \alpha^L, 1) - \tau^*(\alpha^H, \alpha^L, 1).$$

As the incumbent with $\alpha_t = \alpha^L$ distorts her fiscal policy by increasing taxes, this also applies that the constraint is binding for the incumbent with (α^L, α^L). Recall, that $k^*(\alpha^i, \alpha^L, 1) = \alpha^i + \alpha^L + \tau^*(\alpha^i, \alpha^L, 1)$ and $c^*(\alpha^i, \alpha^L, 1) = y - \tau^*(\alpha^i, \alpha^L, 1)$ for $i = L, H$. We can re-write inequality (3.4) as

$$U(c^*(\alpha^H, \alpha^L, 1)) - U(c^*(\alpha^L, \alpha^L, 1)) < \beta \left(V(k^*(\alpha^L, \alpha^L, 1)) - V(k^*(\alpha^H, \alpha^L, 1)) \right). \tag{3.7}$$

As the incumbent increases taxes when distorting, then

$$\beta V'(k^*(\alpha^i, \alpha^L, 1)) < U'(c^*(\alpha^i, \alpha^L, 1)). \tag{3.8}$$

We use this to show that inequality (3.7) cannot hold

$$\begin{aligned} U(c^*(\alpha^H, \alpha^L, 1)) - U(c^*(\alpha^L, \alpha^L, 1)) &\geq U'(c^*(\alpha^H, \alpha^L, 1)) \left(\tau^*(\alpha^L, \alpha^L, 1) - \tau^*(\alpha^H, \alpha^L, 1) \right) \\ &> \beta V'(k^*(\alpha^H, \alpha^L, 1)) \left(\alpha^L + \tau^*(\alpha^L, \alpha^L, 1) - \alpha^H - \tau^*(\alpha^H, \alpha^L, 1) \right) \\ &\geq \beta (V(k^*(\alpha^L, \alpha^L, 1)) - V(k^*(\alpha^H, \alpha^L, 1))) \end{aligned}$$

where the first inequality follows from the fact that $\tau^*(\alpha^L, \alpha^L, 1) > \tau^*(\alpha^H, \alpha^L, 1)$, which implies that $U'(c^*(\alpha^L, \alpha^L, 1)) > U'(c^*(\alpha^H, \alpha^L, 1))$. The second inequality follows from inequality (3.8) and $\alpha^H > \alpha^L$. The concavity and $k^*(\alpha^H, \alpha^L, 1) < k^*(\alpha^L, \alpha^L, 1)$ gives us the last inequality. This contradicts inequality (3.7), and thereby completes this step.

Step 4: We now show, that there exists no separating equilibrium in which the incumbent with the lowest re-election probability calls for a snap election.

We assume that there exists a separating equilibrium in which $\Omega(1,1) > \Omega(0,1)$ and the incumbent calls for a snap election when $\alpha_t = \alpha^L$. We note, that if the incumbent calls for a snap election when $(\alpha_{t-1}, \alpha_t) = (\alpha^i, \alpha^L)$ for $i \in \{L, H\}$, then $\Omega(\rho,1) - \Omega(0,1) > X$ (by Step 1) and

$$(\Omega(0,1) - \Omega(\rho,1) + X)\,(1 - \hat{\pi}^L) + (\Omega(0,0) - \Omega(0,1)) \leq X\left(\hat{\pi}^L - \beta\hat{\pi}^\rho\right)\sum_{s=0}^{\infty}\beta^s\bar{\pi}_s - X\hat{\pi}^L,$$

as the incumbent with (α^i, α^L) prefers to call for a snap election by assumption. This implies that the incumbent will also call for a snap election when $(\alpha_{t-1}, \alpha_t) = (\alpha^j, \alpha^L)$ for $j \neq i$, as the above inequality does not depend on the realisation of α_{t-1}. When the incumbent calls for a snap election whenever $\alpha_t = \alpha^L$, the incumbent with $\alpha_t = \alpha^H$ faces the same constraint whether $\theta_t = 0$ or $\theta_t = 1$ (if $\theta = 0$ is reached in equilibrium). Thus, $\tau^*(\alpha^i, \alpha^H, 0) \leq \tau^*(\alpha^i, \alpha^H, 1)$, $\Omega(0,1) \leq \Omega(0,0)$ and $\Omega(1,1) \leq \Omega(1,0)$, where the inequalities follows from the fact that $\theta_t = 0$ may not be reached in equilibrium, so the value depends on the voter's belief. Furthermore, $\Omega(0,1) < \Omega^L$ implies that $\tau^*(\alpha^L, \alpha^H, 1) < \tau^{FI}(\alpha^L + \alpha^H)$. As she only distorts her policy choices in periods with an election, $e^*(\alpha^L, \alpha^H, 1) = 1$.

As Step 2 implies that the utility loss from distortion is largest when $(\alpha_{t-1}, \alpha_t) = (\alpha^L, \alpha^H)$, then

$$\begin{aligned}\beta\,(\Omega(\rho,1) - \Omega(1,1))(1 - \hat{\pi}^H) + \beta\,(\Omega(1,1) - \Omega(1,0)) + \beta X\left(\hat{\pi}^H - \beta\hat{\pi}^\rho\right)\sum_{s=0}^{\infty}\beta^s\bar{\pi}_s - \beta X \\ \geq W^{FI}(\alpha^H + \alpha^L) - W^*(\alpha^L, \alpha^H, 1) > W^{FI}(2\alpha^H) - W^*(\alpha^H, \alpha^H, 1).\end{aligned} \tag{3.9}$$

Thus, the incumbent with $(\alpha_{t-1}, \alpha_t) = (\alpha^H, \alpha^H)$ also prefers to call for a snap election, i.e. $e^*(\alpha^H, \alpha^H, 1) = 1$. This implies that $\theta = 0$ is never reached in equilibrium.

Furthermore, it implies that the incumbent always calls for a snap election, so $\bar{\pi}_s = \beta^s\,(\hat{\pi}^\rho)^s$, and depending on the off-equilibrium beliefs $\Omega(1,0) - \Omega(1,1) \in [0, \rho(W^{FI}(2\alpha^H) - W^*(\alpha^H, \alpha^H, 1))]$. Hence, we can evaluate the sum on the LHS of inequality (3.9)

$$\sum_{s=0}^{\infty}\beta^s\bar{\pi}_s = \frac{1}{1 - \beta\bar{\pi}^\rho}.$$

Inserting this into inequality (3.9) we derive

$$\begin{aligned}&\beta\left(\Omega(\rho,1)-\Omega(1,1)\right)(1-\hat{\pi}^H)-\beta\left(\Omega(1,0)-\Omega(1,1)\right)-\frac{\beta X(1-\hat{\pi}^H)}{1-\beta\hat{\pi}^\rho}\\&\geq W^{FI}(\alpha^H+\alpha^L)-W^*(\alpha^L,\alpha^H,1)>W^{FI}(2\alpha^H)-W^*(\alpha^H,\alpha^H,1).\end{aligned}$$

As $\Omega(1,1)>\Omega(\rho,1)$, $\Omega(1,1)\geq\Omega(1,0)$ and $\hat{\pi}^H\leq 1$, the LHS is weakly negative. However, $W^{FI}(\alpha^H+\alpha^L)>W^*(\alpha^L,\alpha^H,1)$, so we have a contradiction.

Now, assume that there exists a separating equilibrium in which $\Omega(0,1)>\Omega(1,1)$ and the incumbent calls for a snap election when $(\alpha_{t-1},\alpha_t)=(\alpha^i,\alpha^H)$ for some $i\in\{L,H\}$. By the same arguments as above this implies that $\Omega(\rho,1)-\Omega(1,1)>X$ and the incumbent always calls for a snap election when $\alpha_t=\alpha^H$. Thus, the constraints for separating are the same for $\theta_t=0$ and $\theta_t=1$, and therefore we have $\tau^*(\alpha^i,\alpha^L,0)\geq\tau^*(\alpha^i,\alpha^L,1)$, $\Omega(0,1)\leq\Omega(0,0)$ and $\Omega(1,1)\leq\Omega(1,0)$. Furthermore, $\Omega(1,1)<\Omega^H$ implies that $\tau^*(\alpha^H,\alpha^L,1)>\tau^{FI}(\alpha^H+\alpha^L)$. As she onlu distorts her policy choices in periods with an election, then $e^*(\alpha^H,\alpha^L,1)=1$.

As Step 2 implies that the utility loss from distortion is largest when $(\alpha_{t-1},\alpha_t)=(\alpha^H,\alpha^L)$, then

$$\begin{aligned}&-\beta\left(\Omega(0,1)-\Omega(\rho,1)\right)(1-\hat{\pi}^L)-\beta(\Omega(0,0)-\Omega(0,1))+\beta X\left(\hat{\pi}^L-\beta\hat{\pi}^\rho\right)\sum_{s=0}^{\infty}\beta^s\bar{\pi}_s-\beta X\\&\geq W^{FI}(\alpha^H+\alpha^L)-W^*(\alpha^H,\alpha^L,1)>W^{FI}(2\alpha^L)-W^*(\alpha^L,\alpha^L,1).\end{aligned}\tag{3.10}$$

Thus, the incumbent with $(\alpha_{t-1},\alpha_t)=(\alpha^L,\alpha^L)$ also prefers to call for a snap election, i.e. $e^*(\alpha^L,\alpha^L,1)$. This implies that $\theta=0$ is never reached in equilibrium, and $\bar{\pi}_s=\beta^s\left(\hat{\pi}^\rho\right)^s$. Hence, as above we can re-write inequality (3.10) as follows

$$\begin{aligned}&-\beta\left(\Omega(0,1)-\Omega(\rho,1)\right)-\beta(\Omega(0,0)-\Omega(0,1))-\frac{\beta X(1-\hat{\pi}^L)}{1-\beta\hat{\pi}^\rho}\\&\geq W^{FI}(\alpha^H+\alpha^L)-W^*(\alpha^H,\alpha^L,1)>W^{FI}(2\alpha^L)-W^*(\alpha^L,\alpha^L,1).\end{aligned}$$

As $\Omega(0,1)>\Omega(\rho,1)$, $\Omega(0,1)\geq\Omega(0,0)$ and $\hat{\pi}^L\leq 1$, the LHS is weakly negative. However, $W^{FI}(\alpha^H+\alpha^L>W^*(\alpha^H,\alpha^L,1))$, so we have a contradiction. This concludes the proof of Step 4.

Step 3 and 4 implies that Proposition 3.3.(1) holds, as well as the last statement of Proposition 3.3.

Step 5: Finally, we show that if there exists a separating equilibrium in which the constraint is binding for an incumbent with $(\alpha_{t-1}, \alpha_t) = (\alpha^L, \alpha^H)$ and for an incumbent with $(\alpha_{t-1}, \alpha_t) = (\alpha^H, \alpha^H)$, then the utility loss from separating is largest when $\alpha_{t-1} = \alpha^H$.

Therefore, we assume that such an equilibrium exists. If both constraints are binding in an equilibrium when $\theta_t \in \{0, 1\}$, then the fiscal policy choices of the incumbent with $(\alpha_{t-1}, \alpha_t) = (\alpha^i, \alpha^H)$ for $i \in \{L, H\}$ is implicitly defined by the binding constraint

$$W^{FI}(\alpha^i + \alpha^L) - U(y - \tau^*(\alpha^i, \alpha^H, \theta_t)) - \beta V(\alpha^i + \alpha^L + \tau^*(\alpha^i, \alpha^H, \theta_t)) = A(\theta_t),$$

where

$$A(0) = \beta \left(\hat{\pi}^H - \hat{\pi}^L\right) (\Omega(0,1) - \Omega(\rho,1) + X) + \beta X \left(\hat{\pi}^H - \hat{\pi}^L\right) \sum_{s=1}^{\infty} \beta^s \bar{\pi}_s,$$

and

$$\begin{aligned} A(1) = {} & \beta X \left(\hat{\pi}^H - \beta \hat{\pi}^\rho\right) \sum_{s=0}^{\infty} \beta^s \bar{\pi}_s \\ & - \beta X \hat{\pi}^\rho - \beta \left(1 - \hat{\pi}^H\right) (\Omega(0,1) - \Omega(\rho,1) + X) + \beta \left(\Omega(0,1) - \Omega(0,0)\right). \end{aligned}$$

Note, that $A(\theta_t)$ does not depend on α^i. As V and U are C^1 functions we can view $\tau^*(z, \alpha^H, \theta_t)$ as a continuous function of $z \in \mathbb{R}$. Thus, by the implicit function theorem we derive

$$\frac{\partial \tau^*(z, \alpha^H, \theta_t)}{\partial z} = -\frac{\beta V'(z + \alpha^L + \tau^*(z, \alpha^H, \theta_t))}{\beta V'(z + \alpha^L + \tau^*(z, \alpha^H, \theta_t)) - U'(y - \tau^*(z, \alpha^H, \theta_t))}. \tag{3.11}$$

As the incumbent with $\alpha_t = \alpha^H$ distorts her fiscal policy by decreasing taxes and V and U are concave functions, we have

$$\begin{aligned} U'(y - \tau^*(z, \alpha^H, \theta_t)) < U'(y - \tau^{FI}(z + \alpha^H)) &= \beta V'(z + \alpha^H + \tau^{FI}(z + \alpha^H)) \\ &< \beta V'(z + \alpha^H + \tau^*(z, \alpha^H, \theta_t)) < \beta V'(z + \alpha^L + \tau^*(z, \alpha^H, \theta_t)). \end{aligned} \tag{3.12}$$

This implies that the distorted taxes are higher when $\alpha_{t-1} = \alpha^H$ than when $\alpha_{t-1} = \alpha^L$.

Let $B = \left(\beta V'(z + \alpha^L + \tau^*(z, \alpha^H, \theta_t)) - U'(y - \tau^*(z, \alpha^H, \theta_t))\right)^{-1} > 0$, where the inequality follows from inequality 3.12).

The utility loss from distortion is given by $C(z) = W^{FI}(z + \alpha^H) - W^*(z, \alpha^H, \theta_t)$. By the

envelope theorem, we have

$$\begin{aligned}\frac{\partial C(z)}{\partial z} &= \beta V'(z+\alpha^H+\tau^{FI}(z+\alpha^H)) + U'(y-\tau^*(z,\alpha^H,\theta_t))\frac{\partial \tau^*(z,\alpha^H,\theta_t)}{\partial z} \\ &\quad - \beta V'(z+\alpha^H+\tau^*(z,\alpha^H,\theta_t))\left(1+\frac{\partial \tau^*(z,\alpha^H,\theta_t)}{\partial z}\right) \\ &= BU'(y-\tau^*(z,\alpha^H,\theta_t))\beta\left(V'(z+\alpha^H+\tau^*(z,\alpha^H,\theta_t)) - V'(z+\alpha^H+\tau^{FI}(z+\alpha^H))\right) \\ &\quad + B\beta V'(z+\alpha^L+\tau^*(z,\alpha^H,\theta_t))\left(\beta V'(z+\alpha^H+\tau^{FI}(z+\alpha^H)) - U(y-\tau^*(z,\alpha^H,\theta_t))\right) > 0,\end{aligned}$$

where the equality follows from equation (3.11), and the inequality follows from inequality (3.12) and V and U being strictly increasing functions. Thus, we conclude that the utility loss from distortion is largest when $\alpha_{t-1} = \alpha^H$. This concludes the proof.

3.6.2 Proof of Proposition 3.4

Proof of Proposition 3.4.(1). Assume that there exists a separating equilibrium in which the incumbent never calls for a snap election.

As nobody calls for a snap election in equilibrium, then $\Omega(\mathbb{1}_{i=H}, 1) = \Omega^i$, $\hat{\pi}^i = \pi^i$, $\sum_{s=0}^{\infty}\beta^s\bar{\pi}_s = \frac{1+\beta}{1-\beta^2\pi^\rho}$ and the appeal of snap elections depend on the voters off-equilibrium beliefs. However, we can bound utility of the incumbent with $(\alpha_{t-1},\alpha_t) = (\alpha^i,\alpha^H)$ for $i \in \{L,H\}$ from above by:

$$W^{FI}(\alpha^i+\alpha^H) + \beta\pi^H\Omega^H + \beta(1-\pi^H)\Omega^O + \beta X\pi^H\sum_{s=0}^{\infty}\beta^s\bar{\pi}_s,$$

which corresponds to the case in which the voters attach probability one the incumbent having $\alpha_t = \alpha^H$ when $\tau = \tau^{FI}(\alpha^i+\alpha^H)$, $e=1$ and $\theta=1$. Thus, the incumbent (α^i,α^H) does not benefit from deviating and calling for a snap election if:

$$\begin{aligned}&\Omega^H\pi^H + \Omega^O(1-\pi^H) + X\pi^H\sum_{s=0}^{\infty}\beta^s\bar{\pi}_s \\ &\leq \Omega(1,0) + X + \beta X\pi^\rho\sum_{s=0}^{\infty}\beta^s\bar{\pi}_s,\end{aligned}$$

which can be re-written to the condition in Proposition 3.4.(1). By Proposition 3.3 and Assumption 3.1 the incumbent strictly prefers to call for a snap election when $\alpha_t = \alpha^L$ compared to facing an election today with a re-election probability of π^L. Therefore, there exist off-equilibrium beliefs that sustain such an equilibrium. □

Proof of Proposition 3.4.(2). Assume that there exists a separating equilibrium in which the incumbent calls for a snap election whenever $\alpha_t = \alpha^H$ in which the conditions from Proposition 3.4.(2) holds.

The incumbent with $(\alpha_{t-1}, \alpha_t) = (\alpha^i, \alpha^H)$ for $i = \{L, H\}$ will not deviate and refer from calling a snap election when $\theta_t = 1$ if:

$$\frac{W^{FI}(\alpha^i + \alpha^H) - W^*(\alpha^i, \alpha^H, 1)}{\beta X} + \frac{\Omega(1,1) - \Omega(\rho,1)}{X}(1 - \hat{\pi}^H) - \frac{\Omega(1,1) - \Omega(1,0)}{X}$$
$$\leq \left(\hat{\pi}^H - \beta\hat{\pi}^\rho\right) \sum_{s=0}^{\infty} \beta^s \bar{\pi}_s.$$

As the utility loss from distortion is weakly larger when $\alpha_{t-1} = \alpha^H$ by assumption and $\theta_t = 1$, this implies that if the incumbent with (α^H, α^H) prefers to call for a snap election, then so does the incumbent with (α^L, α^H).

Furthermore, as in Step 3 and 4 in the proof of Proposition 3.1 we can bound the RHS of the above inequality from below by :

$$\frac{2\hat{\pi}^H\left(1 + \beta(\rho\hat{\pi}^H + 1 - \rho)\right) - (2 + \beta)}{2 - \beta^2\left(1 + \beta(\rho\hat{\pi}^H + 1 - \rho)\right)}.$$

Thus, there is no profitable deviation for an incumbent with $\alpha_t = \alpha^H$. As in the proof of Proposition 3.4.(1) there exists off-equilibrium beliefs such that an incumbent with $\alpha_t = \alpha^L$ has no incentive to deviate either. This completes the proof. □