ESSAYS ON THE USE OF INFORMATION IN ESTIMATION AND EFFORT MOTIVATION

A dissertation submitted to the
Department of Economics
at Brown University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

by

Oscar Wahlstrom

Providence, Rhode Island
October 2020

This dissertation by Oscar Wahlstrom is accepted in its present form

by the Department of Economics as satisfying the

dissertation requirements for the degree of Doctor of Philosophy.

Date _____ _____

Roberto Serrano, Advisor

Recommended to the Graduate Council

Date _____ _____

Susanne Schennach, Reader

Date _____ _____

Eric Renault, Reader

Approved by the Graduate Council

Date _____ _____

Andrew G. Campbell, Dean of the Graduate School

VITA

Oscar Wahlstrom was born in Stockholm, Sweden. Before the start of his academic journey he worked as a lumberjack in Sweden and took part in the Swedish Armed Forces. Realizing then that he wanted to pursue the calling of statistics he enrolled at New York University, from which he graduated with a double degree in Mathematics and Economics before coming to Brown University. While working towards his PhD, he was awarded with the Brown Economics department's Graduate Teaching and Third Year Paper Research awards for the 2018-2019 academic years. Oscar received his PhD in Economics in 2020.

ACKNOWLEDGEMENTS

I am forever grateful to my advisor and coordinator Roberto Serrano for his passionate and generous guidance. Not only did he take on the difficult task of advising someone outside of his field, he also did so without hesitation or reservation. In our weekly talks he both encouraged me and helped me think critically about the problems I encountered, enabling me to solve them quickly and efficiently. Without his incredible presence this dissertation would never have happened. Further, I would like to thank my Econometrics advisor Susanne Schennach. The second chapter evolved from an idea which she suggested, and she always provided me with valuable and feedback until its completion. Her continuous support and advice not only allowed the rapid construction of our joint project, but also helped me find confidence in my work. I would also very much like to thank Eric Renault who initially took me under his wing and patiently taught me what it means to do research. The first chapter of this dissertation came out of an idea which formed in relation to the papers he suggested for my field exam. Despite the fact that he moved to Warwick, he continued to guide my thinking remotely. He helped shape my approach and turned me into a researcher capable of dealing with complex questions rigorously. His calm and meticulous influence will stay with me forever.

The third chapter of this thesis would never have happened without the incredible guidance and financial support of Pedro dal Bó. Not only was he instrumental in the design and implementation, but he also helped me and my co-author through the arduous and frustrating IRB approval process. I would also like to direct thanks to Hui-Wen Ng and Oleg Semenov who helped us both in the Z-tree coding and implementation of the experiment. In addition, both Geoffroy de Clippel and Jack Fanning were remarkably helpful in shaping the ideas and questions, something which allowed us to rapidly move beyond the design stage.

Of course I owe deep gratitude to the graduate students who have fought by my side

over the years at Brown, especially Zeky Murra-Anton. We were brothers in arms and he helped me more than he could ever know. He made it fun to work late on problem sets and study for exams while never failing to make me crack a smile. He is a kindred spirit and together we explored the realm of probability theory. I would also like to give a special thanks to Marco Stenborg Petterson. Marco was always there encouraging me through the tough times. He gave me stimulating thoughts and motivated me to work harder on my research, while also helping me whenever I needed it. The three of us pursued knowledge for knowledge's sake, and they both made my stay at Brown so much more than I ever thought it could be. They have both said to me many times how smart I am, but I am nothing compared to them. I sincerely hope our paths will cross again!

I also want to thank my friends and family for their support and encouragement throughout the process. They always had my back and helped me navigate my way forward. Many phone calls have been made across the Atlantic, all of which gave me the energy to keep working towards my goals. Last, but in every imaginable way not least, I want to thank Kristin Petersmann. I was fortunate enough to have someone who had been through everything before me, and knew just how to navigate the difficulties of the program. She has been my role model, with her unwavering discipline and drive which inspired me to keep going when I thought I could not. Be it through laughs and love, presents and adventures, or even a good talking to when I needed it, she has always been there for me. Thank you Kristin, for everything!

PREFACE

This dissertation consists of three self-contained chapters with a focus on the tools which use information in estimation and the effect of dynamic information revelation on effort provision. Chapters 1 and 2 both consider the setting of Moment Condition models, which have gained much popularity in Economics since the introduction of the Generalized Method of Moments (GMM) estimator of Hansen (1982). The Moment Condition itself is a population assumption which, when correctly defined, ties down a relationship between the true value of an unknown parameter and the moments of the Data Generating Process. An alternative estimation method that can lead to substantial improvements over GMM is Minimum Divergence (MD) estimation, which consists of minimally reweighing the data set using a contrast or discrepancy function. This minimal reweighing procedure results in "implied" probabilities, which are the weights that are closest to the uniform weights while ensuring that a sample version of the Moment Condition holds. Chapter 1, "Probabilities implied by Misspecified Moment Conditions", asks the hereto unasked question of whether there is such a thing as a population counterpart to these implied probabilities. In this joint work with Eric Renault, we limit ourselves to the popular subclass of MD called the Empirical Cressie-Read Estimators. This large class contains the most popular estimators in Empirical Likelihood, Exponential Tilting, Continuous Updating GMM, and Minimum Hellinger Distance, where a long standing question has been which estimator is the most desirable. We show that it takes some restrictive conditions on the Moment Conditions and on the contrast function to ensure the existence of the population implied probabilities. In particular, when the moment functions are unbounded the population implied probabilities do not exist for many contrast functions including the most famous one, namely that of Empirical Likelihood. The first consequence of this non-existence, which we address in this chapter, happens when the model is misspecified so that there is no parameter value which satisfies the Moment Constraint assumption. In this setting

we show that there is no way to define and conduct inference about a pseudo-true value, something which the literature on robustness to misspecification has relied on previously.

While Chapter 1 shows that Empirical Likelihood is not attractive from a misspecification standpoint, it still retains the best higher order efficiency properties in the class of MD estimators. In fact, as a direct corollary of the results in Chapter 1 there are no estimators in the popular Empirical Cressie-Read class which are both higher order efficient and are equipped to handle misspecification. Chapter 2, "Bounded Tilting Estimation", moves beyond the Empirical Cressie-Read and proposes a new class of estimators which simultaneously has both of the desirable properties. In this joint work with Susanne Schennach, we propose the Higher Order Efficient Bounded Tilting Estimator which is also a subclass of the MD estimators but defined using a discrepancy function that satisfies some intricate properties. The first such property is that the resulting tilting function, a mapping which uses the data points and parameter value to compute the implied probabilities, must be bounded which immediately grants guaranteed existence of the population implied probabilities. Secondly, by further limiting the discrepancy function to satisfy some differentiability properties, it is also higher order efficient in the sense of the second order mean squared error of its bias corrected version.

Chapter 3, "Responses to Information Obfuscation in the Laboratory", switches focus away from Econometric theory towards exploring the use of information in effort motivation in a laboratory setting. In this joint work with Kristin Petersmann, we explore situations where the principal cannot motivate the agent with money. Instead, information about the unknown difficulty of a task can be revealed over time by the principal to incentivize the agent to provide effort. Ely and Szydlowski (2020) derived a theoretically optimal information disclosure policy for this setup. It consists of the principal sending a signal about the true task difficulty *after* the agent has exerted a certain amount of effort. We conduct an experiment which first explores whether subjects act as predicted by the theory in responding to the choice of information structure and further tests whether reci-

procity plays a role in their responses. We find significant deviations from the theoretically optimal responses, which cause an information disclosure policy that delays information revelation just as the theoretically optimal one does but avoids the information obfuscation to perform better overall. Further, unlike in the labor market setting of Charness (2004), we do not observe the presence of reciprocity in the subjects' decisions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

PROBABILITIES IMPLIED BY MISSPECIFIED MOMENT CONDITIONS

1.1 Introduction

We consider throughout a H-dimensional random vector Y_θ indexed by some unknown parameter $\theta \in \Theta$. Y_θ is a known function $K(Z, \theta)$ of θ and of some primitive random vector Z, and is hence endowed with a probability measure P_θ which depends on θ. The parameter of interest θ may be finite or infinite dimensional, both of which have received much attention in the literature (see for example Hansen (1982) for the finite dimensional case, and Ai and Chen (2003), Ai and Chen (2007), Darolles et al. (2011), or Otsu (2007) for the infinite dimensional case). We are interested in inference on θ based on the moment conditions:

$$E\left[Y_\theta\right] = 0 \tag{1.1}$$

The moment model is said to be well-specified (resp. misspecified) if there exists (resp. there does not exist) a parameter value $\theta \in \Theta$ for which the conditions (1.1) are fulfilled. Irrespective of whether the model is misspecified or not, the recent literature on inference by minimum contrast (also referred to as minimum divergence, see for example Corcoran (1998)) starts by looking for a change of measure $M_\theta \geq 0$ for each θ, which is the solution of:

$$\begin{aligned} & \min_M E\left[\phi(M)\right] \\ s.t.\, E[M] &= 1 \\ E\left[MY_\theta\right] &= 0 \end{aligned} \tag{1.2}$$

where ϕ is a given contrast function and the expectations are taken with respect to P_θ. ϕ being referred to as a contrast function simply means that it is a strictly convex function defined on $\mathbb{R}_+^*$ such that:[1]

$$\phi(1) = 0$$

In practice, the statistician will minimize a sample counterpart of (1.2) for each given value of θ. From a sample $\{Z_i\}_{i=1}^n$, the solution of the sample counterpart to (1.2) will deliver implied probabilities $\{q_{i,n}^\gamma(\theta)\}_{i=1}^n$ that depend on the data only through the vector $(Y_{i,\theta})_{1\leq i\leq n} = [K(Z_i,\theta)]_{1\leq i\leq n}$. Even more importantly, it can be shown that in terms of first order asymptotics, it is immaterial to replace each implied probability $q_{i,n}^\gamma(\theta)$ by an infeasible one $q_{i,n}^{\gamma*}(\theta)$ that depends on the data only through $Y_{i,\theta}$ for the same observation number i. $q_{i,n}^\gamma(\theta)$ depends on $Y_{j,\theta}, j \neq i$ only through sample counterparts to population moments of Y_θ and by replacing these sample counterparts by the actual population moments we get what we refer to as the infeasible implied probability $q_{i,n}^{\gamma*}(\theta)$. Intuitively then, since there is no first order asymptotic difference between sample moments and population moments under some i.i.d assumption, there is no first order asymptotic difference between the two implied probabilities. Note however that this point of view overlooks the possibility to pre-average the consecutive values $Y_{i+h,\theta}, h = -\varpi_n, ... -1, 0, 1, ...\varpi_n$ with a convenient bandwidth ϖ_n to take care of serial dependence in moment functions.

The fact that $q_{i,n}^{\gamma*}(\theta)$ depends on the data $Z_j, j = 1, ..., n$, only through $Y_{i,\theta}$ justifies that we consider in the population minimization problem (1.2) only changes of measure $M = M(Y_\theta)$ that depend on the state of nature only through the value of the random moment function Y_θ. The rationale for this minimization is an obvious implication of Jensen inequality (jointly with $\phi(1) = 0$), telling us that:

On the one hand:

[1]This condition ensures that no change of measure at a point y, i.e. $M(y) = 1$, leads to no contrast between the measures at y.

$$E[M] = 1 \Longrightarrow E\left[\phi(M)\right] \geq 0$$

and on the other hand when not only $E[M] = 1$ but also $E\left[MY_\theta\right] = 0$, then:

$$E\left[\phi(M)\right] = 0 \Longleftrightarrow E\left[Y_\theta\right] = 0$$

In particular, the value of objective function of the minimization problem (1.2) is zero if and only if $E\left[Y_\theta\right] = 0$. In this case, the minimum is reached at only one (up to an almost sure equality) change of measure M_θ which is identical to the constant 1 and hence does not change the measure at all.

More generally, the change of measure M_θ defines the probability density function of a probability measure Q_θ with respect to the probability measure P_θ of the random vector Y_θ:

$$M_\theta(y) \quad = \quad \frac{dQ_\theta}{dP_\theta}(y)$$

When applied to the probability distribution P_θ the change of measure (aptly for its name) changes P_θ into Q_θ, since $M_\theta(y)dP_\theta(y) = dQ_\theta(y)$. This in particular means that:

$$E\left[M_\theta g(Y_\theta)\right] \quad = \quad \int M_\theta(y)g(y)dP_\theta(y) = \int g(y)dQ_\theta(y)$$

For obvious reasons, it is natural to dub the probability measure $Q_\theta = M_\theta P_\theta$ the "population implied probabilities". For any transformation $g(Y_\theta)$ of Y_θ, the weighted average of observations $\{g(Y_{i,\theta})\}_{i=1}^n$ computed with respective weights $q_{i,n}^\gamma(\theta)$:

$$\sum_{i=1}^{n} q_{i,n}^\gamma(\theta)g(Y_{i,\theta})$$

is the sample analog of the population expectation $E\left[M_\theta g\left(Y_\theta\right)\right]$.

The focus and interest of this chapter is on the existence and uniqueness of the population implied probabilities, and hence naturally on the change of measure M_θ solution of (1.2), in the case when Y_θ has a non-zero mean. While clearly relevant for misspecified moment models, note that this issue is relevant even for the study of well-specified moment models since then the moment vector $E\left[Y_\theta\right]$ is different from zero except for some true unknown value(s) θ^0 of the parameters θ of interest.

When it is well-defined, a possible use of this change of measure is to allow inference on θ through the minimization of the sample counterpart of the population profile criterion, defined by the following optimization problem:

$$\min_{\theta\in\Theta} E\left[\phi(M_\theta)\right] \tag{1.3}$$

In the case of a well-specified (resp. misspecified) moment model, the minimization of the population profile criterion (1.3) is supposed to characterize the true unknown value θ^0 (resp. the pseudo-true value θ^*) of θ. The sample counterpart of this minimization is common practice in the Minimum Empirical Discrepancy and GEL literature. An important message of the present chapter is that, as far as population criterions are concerned, the existence of the solution M_θ to the minimization (1.2) is far from being granted. Of course, the implications of this population issue would differ depending on whether the moment model is well-specified or misspecified.

In the case of a well-specified moment model that identifies the true unknown value θ^0, θ^0 defines the only change of measure $M = M_{\theta^0}$ such that $E\left[\phi(M)\right] = 0$. Since $E\left[\phi(M)\right] > 0$ for any other change of measure consistent with the constraints of the minimization problem (1.2), one can consider by abuse of language that M_{θ^0} solves the minimization program (1.3), even though M_θ may not be defined for some values $\theta \in \Theta$. However, this remark may lead one to question the interpretation of the implied probabilites $q_{i,n}^{\gamma}(\theta)$ for

the $\theta \neq \theta^0$ for which the population counterpart does not exist.

In the case of a misspecified moment model, this issue is even more harmful since then it is possible that M_θ does not exist for all $\theta \in \Theta$, in which case a pseudo-true value θ^* (about which inference would be sensible) is impossible to define in the classical sense.

While these specific issues will be discussed in more detail in the conclusion, the first task of this chapter is to state general conditions under which population implied probabilities may or may not exist for values of θ such that $E(Y_\theta) \neq 0$. In a nutshell, following the seminal work of Csiszár (1995), sufficient conditions for existence of implied probabilities can be stated, either through assumptions on the moment functions (by assuming that the support of the random vector Y_θ is bounded), or through assumptions on the contrast function (by assuming that ϕ is differentiable with a derivative ϕ' such that $\lim_{m=\infty} \phi'(m) = +\infty$) . We will then show that these sufficient conditions are close to being necessary by setting a special focus on the Cressie-Read family of constrasts ϕ_γ, indexed by $\gamma \in \mathbb{R}$, and defined by:

$$\phi_\gamma(M) = \frac{M^{\gamma+1} - 1}{\gamma(\gamma+1)}, \forall \gamma \neq 0 \tag{1.4}$$

knowing that the above definition for $\gamma = -1$ must be understood as a limit case:

$$\phi_{-1}(M) = -\log(M)$$

while, following a common convention:

$$\phi_0(M) = M\log(M)$$

We note that:

$$\lim_{m=\infty} \phi'_\gamma(m) = +\infty \Longleftrightarrow \gamma \geq 0$$

When this condition is violated and the moment condition is non-zero, that is $\gamma < 0$ and $E[Y_\theta] \neq 0$, we can show that it takes a very small departure from a bounded moment function for the optimization problem (1.2) to not have a solution. Namely, it takes an absolutely continuous variable Y_θ with bounded strictly positive density in a neighborhood of the line $\alpha E[Y_\theta]$, $\alpha > 0$ to be able to build a sequence $\{M^{(j)}\}_{j=1}^{\infty}$ of changes of measure consistent with the constraints of (1.2) such that:

$$\lim_{j=\infty} E\left[\phi_\gamma(M^{(j)})\right] = 0 \tag{1.5}$$

The construction of this sequence takes only a minor extension of a proof initially proposed by Hansen et al. (2016) (HHM hereafter). While HHM's initial result was for the case of Empirical Likelihood (EL), that is $\phi_{-1}(M) = -\log(M)$, we extend the result to what we will dub *Empirical-Likelihood-like* (EL-like) constrast functions, which are the ϕ_γ indexed by $\gamma < 0$. Note that for any M consistent with the constraints of (1.2):

$$E[Y_\theta] \neq 0 \Longrightarrow E[\phi_\gamma(M)] > 0$$

so that the limit result (1.5) implies that a solution M_θ for the minimization (1.2) does not exist for the EL-like contrasts $\phi = \phi_\gamma, \gamma < 0$.

By contrast, when $\gamma \geq 0$ a solution M_θ exists for the minimization (1.2) even if Y_θ is unbounded under regularity condition due to Csiszár (1995). By extension of the case of Euclidean Empirical Likelihood ($\gamma = 1$), we will dub *Chi-Square-like* (χ^2-like) all the constrast functions ϕ_γ indexed by $\gamma \geq 0$.

The rest of this chapter is organized as follows. Section 2 explores and characterizes the conditions for existence using the aforementioned work of Csiszár, Section 3 then gives the non-existence result for the EL-like contrast functions, and Section 4 concludes by examining closer the implications of this non-existence.

1.2 Conditions for Existence of Population Implied Probabilities

1.2.1 The case of bounded variables

For a given value $\theta \in \Theta$ we consider first the case when $||Y_\theta||_\infty < \infty$. With some slight abuse of notation the boundedness of the H components $Y_{j,\theta}$ of Y_θ allows us to consider $2H$ non-negative random variables a_j:[2]

$$a_j(Y) = L - Y_{j,\theta}, a_{j+H}(Y) = Y_{j,\theta} - l, j = 1, ..., H$$

where it is assumed that we have with probability one, for all $j = 1, .., H$:

$$l \leq Y_{j,\theta} \leq L$$

Note that, by considering a given value of θ, we simplify notation by not making explicit the dependence of the functions $a_j(Y), j = 1, ..., 2H$ and of l and L on θ. Then, for any probability density function $M(y)$ w.r.t. P_θ we can characterize the constraint $E[MY_\theta] = 0$ by the following system of $2H$ inequalities:

$$\begin{aligned} \int a_j(y)M(y)dP_\theta(y) &\leq L, \forall j = 1, ..., H \\ \int a_j(y)M(y)dP_\theta(y) &\leq -l, \forall j = H+1, ..., 2H \end{aligned} \tag{1.6}$$

In order to apply the results of Csiszár (1995), we will maintain the assumption:

Assumption A1: The probability distribution P_θ of the random vector Y_θ is absolutely

[2]In order for this not to be confused with the realizations $Y_{i,\theta}$ discussed in the introduction we use j to denote which member of the H-dimensional vector Y_θ we are referring to

continuous with respect to some σ-finite measure λ:

$$\frac{dP_\theta}{d\lambda}(y) = h_\theta(y)$$

and $h_\theta(y) > 0$ $\lambda-$almost everywhere.

Note that the strict positivity of $h_\theta(y)$ $\lambda-$almost everywhere is hardly restrictive since by definition:

$$[h_\theta(y) = 0, \forall y \in B] \Longrightarrow P_\theta(B) = 0$$

and then, the dominating measure λ can always be chosen such that $\lambda(B) = 0$. In this context, Cziszár studies the linear inverse problem (1.6) by looking for a probability density function $s(y)$ with respect to the measure λ solution of a minimization problem:

$$\begin{aligned} & \min_s \int h_\theta(y) f\left(\frac{s(y)}{h_\theta(y)}\right) d\lambda(y) \\ s.t. \int a_j(y)s(y)d\lambda(y) &\leq L, \forall j = 1, ..., H \\ \int a_j(y)s(y)d\lambda(y) &\leq -l, \forall j = H+1, ..., 2H \end{aligned} \tag{1.7}$$

where f is a given differentiable strictly convex function on $\mathbb{R}_*^+$, satisfying:

$$f(1) = f'(1) = 0$$

The objective funtion of (1.7) is well-defined precisely because $h_\theta(y) > 0$ $\lambda-$almost everywhere. Note that if we consider a contrast function ϕ as in the introduction that is differentiable, we get a well-suited function f by considering:

$$f(u) = \phi(u) - \phi'(1)\,[u - 1]$$

Note also that if M stands for the probability density with respect to P_θ of some

probability measure Q on $\mathbb{R}^H$, we can define $s = Mh_\theta$ and check that it is indeed a probability density function with respect to the measure λ:

$$s(y) = M(y)\, h_\theta(y) = \frac{dQ}{dP_\theta}(y)\frac{dP_\theta}{d\lambda}(y) = \frac{dQ}{d\lambda}(y) \tag{1.8}$$

The fact that Q is then a probability measure gives us the second constraint $E[M] = \int M(y)h_\theta(y)d\lambda(y) = \int dQ(y) = 1$ trivially. Note then that by rewriting (1.7) with the notations of (1.8) (and using M as the argument for the minimization), we get

$$\begin{aligned} & \min_M \int f\left[M(y)\right] dP_\theta(y) \\ s.t. \int a_j(y)M(y)dP_\theta(y) &\leq L, \forall j = 1, ..., H \\ \int a_j(y)M(y)dP_\theta(y) &\leq -l, \forall j = H+1, ..., 2H \end{aligned}$$

which can be rewritten:

$$\begin{aligned} & \min_M E\left[f\left[M\left(Y_\theta\right)\right]\right] \\ s.t. E\left[M\left(Y_\theta\right)Y_\theta\right] &= 0 \end{aligned}$$

Moreover, it is worth noting that:

$$E\left[f\left[M\left(Y_\theta\right)\right]\right] = E\left[\phi\left[M\left(Y_\theta\right)\right]\right] - \phi'(1)\left[E\left[M\left(Y_\theta\right)\right] - 1\right] = E\left[\phi\left[M\left(Y_\theta\right)\right]\right]$$

since by definition, as mentioned before, $E[M(Y_\theta)] = 1$. In other words, the minimization problem (1.7) is nothing but the minimization problem of interest (1.2) with the change of variable $M \to s$.

Regarding the minimization problem (1.7), Theorem 3(i), p177, in Csiszár (1995) tells us that, thanks to the non-negativity of the random variables $a_j(Y_\theta), j = 1, ..., 2H$ and to Assumption A1, a solution s_θ to the problem (1.7) (the so-called D-projection problem

in Cziszar's terminology) always exists. Then, from the above discussion, we do have a solution:

$$M_\theta(y) = \frac{s_\theta(y)}{h_\theta(y)}, \lambda - ae$$

to our problem of interest (1.2). The function s_θ is the D-projection of h_θ on the set of functions s defined by inequalities (1.6) (with $M(y)$ replaced by $s(y)/h_\theta(y)$).

Remark: The boundedness assumption is popular in the model selection literature, see for example Chen et al. (2007), but is mostly referred to ensure $\sqrt{n}$ convergence of the solution to the sample counterpart to our minimization problem over θ (1.3). The message of the exposition above is then that such an assumption also could, and in some situations perhaps should, be made in order to guarantee the reweighting problem (1.2) actually has a solution in the population when the model is misspecified so that, under some regularity conditions, the problem (1.3) also has a solution in the population.

1.2.2 The general case

As discussed in the introduction, unboundedness may be harmful to the minimization problem (1.2). To better understand why this is and to explore what conditions on contrast functions may provide a hedge against it, it is worth looking at the first order conditions of the minimization problem (1.2) in the context of assumption 1. The Lagrangian function of (1.2) is:

$$\mathcal{L} = \int \phi\left[M(y)\right] h_\theta(y) d\lambda(y) - a\left\{\int M(y) h_\theta(y) d\lambda(y) - 1\right\} - b'\left\{\int M(y) y h_\theta(y) d\lambda(y)\right\}$$

where $a \in \mathbb{R}$ and $b \in \mathbb{R}^H$ are Lagrange multipliers corresponding to the constraints $\mathbb{E}[M] = 1$ and $\mathbb{E}[MY_\theta] = 0$ respectively. Under mild regularity conditions including a differentiable contrast function, the first order conditions become (for $\lambda - a.e.$ value of y)

after differentiation w.r.t. $M(y)$:

$$\phi' \left[M(y) \right] h_\theta(y) - a h_\theta(y) - b' y h_\theta(y) = 0 \tag{1.9}$$

By right-multiplying by $M(y)$ (resp. $M(y)y'$), integrating w.r.t. y, and using the constraints of (1.2), we get one equation (resp. H equations) to determine the Lagrange multipliers a and b respectively:

$$\begin{aligned} a^* &= E\left[M(Y_\theta) \phi' \left[M(Y_\theta) \right] \right] \\ E\left[M(Y_\theta) Y_\theta Y_\theta' \right] b^* &= E\left[M(Y_\theta) Y_\theta \phi' \left[M(Y_\theta) \right] \right] \end{aligned}$$

By plugging these values of a and b into (1.9), we get the optimal value of $M(y)$ for all y (up to $\lambda - a.e.$ equality) by inverting the, thanks to strict convexity, strictly increasing function ϕ'. Since Assumption A1 guarantees the density to be non-zero, we can write:

$$\phi' \left[M(y) \right] = a^* + b^{*\prime} y, \lambda - a.e.$$

In particular if a solution M_θ exists, we will have almost surely:

$$\phi' \left[M_\theta(Y_\theta) \right] = a^* + b^{*\prime} Y_\theta \tag{1.10}$$

The identity (1.10) displays clearly the issue we are facing for existence of a solution M_θ. If the random variable Y_θ is not bounded, the linear function $[a^* + b^{*\prime} Y_\theta]$ is not bounded either. Since the function ϕ' is strictly increasing, the divergence of Y_θ must be coupled with a divergence of the density function M_θ, leading to the divergence of $\phi' \left[M_\theta(Y_\theta) \right]$ thanks our next maintained assumption A2:

Assumption A2:

$$\lim_{m=\infty} \phi'(m) = +\infty$$

In the context of the Cressie-Read family (1.4) of contrasts:

$$\begin{aligned}\phi'_\gamma(m) &= \frac{m^\gamma}{\gamma}, \forall \gamma \neq 0 \\ \phi'_0(m) &= \log(m) + 1\end{aligned}$$

we note that assumption A2 is fulfilled if and only if $\gamma \geq 0$, as opposed to the EL-like contrasts ($\gamma < 0$) where:

$$\lim_{m=\infty} \phi'_\gamma(m) = 0$$

While assumption A2 appears to be necessary for the existence of M_θ in case of an unbounded variable Y_θ (as confirmed by the pretty general construction in the next section of a counter-example for all EL-like contrasts), we can again use Csiszár (1995) to show to what extent it is sufficient.

If we want to relax the boundedness assumption on Y_θ, we can simply consider the system (1.6) of inequalities with arbitrary values of numbers l and L (that are not bounds anymore), for instance $l = L = 0$, and variables $a_j(Y), j = 1, .., 2H$, which are not assumed anymore to be non-negative. As in the former subsection, we still note the equivalence betwen Cziszar's projection problem (1.7) and our problem of interest (1.2), through the change of variable $M \longrightarrow s$.

Regarding the minimization problem (1.7), Theorem 3(iii), p177, in Csiszár (1995) tells us that, thanks to Assumptions A1 and A2, and in spite of the fact that the functions $a_j(Y), j = 1, .., 2H$ may take both positive and negative values, a solution s_θ to the problem (1.7) always exists as soon as:

$$\int f^* \left[\alpha a_j^-(y)\right] h_\theta(y) d\lambda(y) < \infty, \forall \alpha > 0, \forall j = 1, ..., 2H$$

where:

$$a_j^-(y) = \max\left(0, -a_j(y)\right)$$

and f^* denotes the convex conjugate of f:

$$f^*(v) = \sup_u [uv - f(u)]$$

As Csiszár reminds us (see formula (3.3) p177), our assumption A2 allows us to characterize the convex conjugate of $f(u) = \phi(u) - \phi'(1)[u-1]$ as follows:

$$f^*(v) = \int_0^v (f')^{-1}(z)dz = \int_0^v (\phi')^{-1}[z + \phi'(1)]\,dz$$

To fit our optimization problem (1.2) into the context we choose our functions:

$$a_j(Y) = -Y_{j,\theta}, a_{j+H}(Y) = Y_{j,\theta}, j = 1, ..., H$$

Which then, when applied to the assumption made by Csiszár, becomes:

Assumption A3: For all $\alpha > 0$ and all $j = 1, ..., H$:

$$E\left\{f^*\left[\alpha Y_{j,\theta}^+\right]\right\} < \infty, E\left\{f^*\left[\alpha Y_{j,\theta}^-\right]\right\} < \infty$$

where:

$$\begin{aligned} y^+ &= \max(y, 0), y^- = \max(-y, 0) \\ f^*(v) &= \int_0^v (\phi')^{-1}[z + \phi'(1)]\,dz \end{aligned}$$

Then, from the above discussion, under Assumptions A1, A2, and A3 a solution in the form:

$$M_\theta(y) = \frac{s_\theta(y)}{h_\theta(y)}, \lambda - ae$$

exists to our problem of interest (1.2).

This result ensures very generally the existence of implied probabilities for any χ^2-like Cressie Read contrast ϕ_γ,$\gamma \geq 0$, since we can show:

Lemma 1: For the Cressie-Read contrast function ϕ_γ with $\gamma \geq 0$, a necessary and sufficient condition for assumption A3 with $\phi = \phi_\gamma$ is:

For $\gamma > 0, |Y_{j,\theta}|^{\frac{\gamma+1}{\gamma}}$ is integrable for all $j = 1, ..., H$.

For $\gamma = 0$, Y_θ has a finite Laplace transform $E\left[\exp\left(t'Y_{j,\theta}\right)\right]$ for all $t' \in \mathbb{R}^h$ for all $j = 1, ..., H$.

Not surprisingly, the smaller the index γ, the more restrictive is the needed integrability assumptions on Y_θ for existence of implied probabilities. The condition for $\gamma = 0$ is tantamount to assume the integrability at any order, which is as expected the limit case (when $\gamma \longrightarrow 0$) of the assumption needed in the case $\gamma > 0$. However, it is worth noting that the necessary and sufficient condition put forward by lemma 1 is very natural. To see that, we first note that when using the contrast function ϕ_γ to solve (1.2), we work with changes of measure $M \geq 0$ such that $\phi_\gamma(M)$ is integrable, meaning with standard notations that $M \in L^{\gamma+1}$ when $\gamma \geq 0$. Thus, we want that:

$$M \in L^{\gamma+1} \Longrightarrow MY_{j,\theta} \in L^1, \forall j = 1, .., H$$

in order to be able to impose the constraint $E\left[MY_\theta\right] = 0$. By virtue of Holder's inequality, this assumption will be fullfilled if:

$$Y_{j,\theta} \in L^p, \forall j = 1, .., H$$

such that:

$$\frac{1}{p} + \frac{1}{\gamma + 1} = 1$$

that is:

$$p = \frac{\gamma + 1}{\gamma}$$

that is nothing but the condition put forward for lemma 1. For instance, with Euclidean Empirical Likelihood ($\gamma = 1$), we need to use changes of measure M with finite variance and the corresponding moment functions, components of Y_θ, must have finite variance as well.

1.3 Non-Existence of Population Implied Probabilities for EL-like Contrasts

While the previous section provided good news in the shape of sufficient conditions for existence, as already announced in the introduction we can prove the non-existence of population implied probabilities by building a sequence of changes of measure consistent with the constraints of (1.2) such that:

$$\lim_{j=\infty} E\left[\phi_\gamma(M^{(j)})\right] = 0 \tag{1.11}$$

Our first remark is that as long as the minimization (1.2) with $\phi = \phi_\gamma$ amounts to the maximization of $E\left(M^{\gamma+1}\right)$ (meaning $-1 < \gamma < 0$), any sequence $\{M^{(j)}\}_{j=1}^{\infty}$ of changes of measure leading to the zero-limit in (1.5) with $\gamma = -1$, also leads to a zero limit with $\phi = \phi_\gamma$ for any γ in $]-1, 0[$:

Lemma 2: If $\{M^{(j)}\}_{j=1}^{\infty}$ is a sequence of random variables of unit expectation such that:

$$\lim_{j=\infty} E\left[\log(M^{(j)})\right] = 0 \tag{1.12}$$

Then:

$$\lim_{j=\infty} E\left[\phi_\gamma(M^{(j)})\right] = 0, \forall \gamma \in]-1, 0[$$

Fix a $\theta \in \Theta$ such that $E[Y_\theta] \neq 0$. In order to build a sequence $\{M^{(j)}\}_{j=1}^{\infty}$ which fulfills the constraints of (1.2) and is conformable to (1.12), we will slightly extend a proof by HHM.

First, for any $\varepsilon > 0$ and $\alpha > 0$, $G_\theta^\varepsilon(\alpha)$ stands for the closed ball of center $[-\alpha E(Y_\theta)]$ and radius ε. We will then maintain the following assumption:

Assumption A4:

(i) The variable Y_θ is absolutely continuous (with respect to the Lebesgue measure) with a continuous probability density function h_θ.

(ii) There exists $\bar{\varepsilon} > 0$ and $B > 0$ such that, for all $\alpha > 0$, h_θ is strictly positive on $G_\theta^{\bar{\varepsilon}}(\alpha)$ and:

$$\sup\{h_\theta(y); y \in G_\theta^{\bar{\varepsilon}}(\alpha)\} \leq B$$

(iii) There exists $\bar{\alpha} > 0$ such that:

$$\sup_{\alpha > \bar{\alpha}} \{h_\theta(y); y \in G_\theta^{\bar{\varepsilon}}(\alpha)\} \leq 1$$

It is worthwhile to emphasize the difference between Assumptions A4 (ii) and (iii). On the one hand, Assumption A4(ii) only maintains that the continuous function h_θ is bounded on any compact set $G_\theta^{\bar{\varepsilon}}(\alpha)$ indexed by $\alpha > 0$, and since all these compact sets are balls of the same radius $\bar{\varepsilon}$, we may assume that the bound is uniform over $\alpha > 0$. On the other hand, since the function h_θ is integrable, it goes to zero at infinity in any direction. Thus, when the center $[-\alpha E(Y_\theta)]$ of the ball $G_\theta^{\bar{\varepsilon}}(\alpha)$ is large enough, we simply make sure that h_θ is sufficiently small in the ball. Hence Assumption A4(iii).

The intuition behind HMM's construction of the requested sequence $\{M^{(j)}\}_{j=1}^{\infty}$ lies in the observation we made in the previous section that there is a problem when Y_θ takes on values that become arbitrarily large and $\gamma < 0$ (for which the first derivative of the contrast function does not go to ∞). The key idea is then to consider for some $\pi \in]0,1[$ and some $l > 0$ a random variable:

$$M = 1 - \pi + \frac{\pi}{l}\frac{1_G(Y_\theta)}{h(Y_\theta)} \tag{1.13}$$

where 1_G stands for the indicator function of a set G:

$$\begin{aligned} 1_G(y) &= 1 \; if \; y \in G \\ 1_G(y) &= 0 \; otherwise \end{aligned}$$

and $G = G_\theta^\varepsilon(\alpha)$ for a convenient choice of ε and α. This is tantamount to the change of measure M taking a mass π and distributing it into the neighborhood G while reweighting the neighborhood so that the resulting distribution is uniform on G. We note that the numbers l, ε, α can be easily chosen so that M in (1.13) fulfills the constraints of (1.2):

First, $E(M) = 1$ if and only if l is the Lebesque measure of G,

Second, with this choice of l, $E(MY_\theta) = 0$ if and only if:

$$\alpha = \frac{1-\pi}{\pi}$$

To see that, note that by considering the expectation of the uniform distribution on the ball with center $[-\alpha E(Y_\theta)]$, we have:

$$\int_{G_\theta^\varepsilon(\alpha)} \frac{y}{l} dy = -\alpha E(Y_\theta)$$

Note that at this stage ε is not subject to any constraint except that we assume $\varepsilon \leq \bar{\varepsilon}$

to be sure that $h(.)$ is strictly positive on $G_\theta^\varepsilon(\alpha)$. For the next step we will however allow ϵ to vary with j, in other words to consider ϵ_j.

HHM build the requested sequence $M^{(j)}$ by applying (1.13) with an arbitrary sequence $\pi = \pi_j$ going to zero when j goes to infinity. Then, we obviously have a sequence $M^{(j)}$ converging to 1 in probability if:

$$\lim_{j=\infty} \Pr\left[Y_\theta \in G_\theta^{\varepsilon_j}(\alpha_j)\right] = 0, \alpha_j = \frac{1-\pi_j}{\pi_j}$$

This limit will be warranted by choosing the sequence of radius ε_j going itself to zero when j goes to infinity. The challenge will then be to choose this sequence in order to also ensure the desired limit:

$$\lim_{j=\infty} E\left[\phi_\gamma(M^{(j)})\right] = 0, \forall \gamma < 0$$

It is worth keeping in mind that this construction tightly relies on the fact that the moment function Y_θ is not bounded since it has a strictly positive density on the balls $G_\theta^{\varepsilon_j}(\alpha_j)$ whose center $[-\alpha_j E(Y_\theta)]$ drifts to infinity when j goes to infinity.

We can then show:

Theorem 1: For any sequence $\pi_j \in]0,1[$ going to zero and:

$$\begin{aligned} \alpha_j &= \frac{1-\pi_j}{\pi_j} \\ M_j(\theta) &= 1 - \pi_j + \frac{\pi_j}{l_j(\theta)} \frac{1_{G_j(\theta)}(Y_\theta)}{h(Y_\theta)} \\ G_j(\theta) &= G_\theta^{\varepsilon_j}(\alpha_j) \end{aligned}$$

where $l_j(\theta)$ is the Lebesgue measure of $G_\theta^{\varepsilon_j}(\alpha_j)$, there exists a sequence ε_j on $]0, \bar{\varepsilon}[$ such that:

$$\lim_{j=\infty} E\left[\phi_\gamma(M^{(j)})\right] = 0, \forall \gamma < 0$$

The proof of theorem 1 is given in the Appendix. For $\gamma = -1$, it closely follows the proof proposed by HHM. The proof is even simpler for $\gamma < -1$ since it does not require a specific choice of the sequence ε_j going to zero. For $\gamma \in]-1, 0[$, there is no additional proof needed thanks to lemma 1.

A key implication of our Theorem 1 is then simply that there can be no such thing as population implied probabilities for an unbounded continuous random variable when using EL-like contrasts for all θ which do not satisfy the moment condition (1.1).

1.4 Conclusion

Given the results in sections 2 and 3, we must discuss their implications for researchers. Combining the non-existence and existence results for EL-like contrasts, we see that unless the moment function Y_θ is bounded there is no solution to the population reweighting problem for that θ when $E[Y_\theta] \neq 0$. For misspecified models ($E[Y_\theta] \neq 0 \; \forall \theta \in \Theta$) the population profile criterion (1.3) can hence no longer have a solution. This in particular means that the standard robustness analysis technique of analyzing convergence of an estimator $\hat{\theta}$ towards a pseudo-true value (the solution to (1.3)) no longer works since the pseudo-true value defined in that way no longer exists. While this does not preclude convergence of the estimator, it continues and completes the analysis of Schennach (2007) since indeed this also means that $\sqrt{n}$ convergence towards the pseudo-true value is impossible. Hence in the language of the misspecification literature, any estimator which uses an EL-like contrast function will not be robust to (global) misspecification.

This problem is solved for the χ^2-like contrasts by the sufficient condition provided in section 2. This together with some regularity conditions on the function $K(z, \theta)$ gives existence of a pseudo-true value around which inference can be made. Given this result it seems natural to assume that a pseudo-true value will also exist if one solves the reweighting problem with the χ^2-like contrasts and the optimization over θ with a EL-like

contrast, as is done in Schennach's ETEL estimator and the ETHD estimator of Antoine and Dovonon (2018). As an avenue for future research, this however also implies that any estimator using such a combination method should be able to guarantee robustness to misspecification given that one can use a standard "uniform convergence of optimization problem" proof approach for consistency. This mixing of contrasts is also relevant in testing where Chaudhuri and Renault (2017) have proposed a novel way to use more than one contrast. To allude to another future avenue of research, in this setting it also appears massively important that the implied probabilities behave nicely under misspecification, since the alternative is by definition always misspecified under the null, if one is looking to guarantee equivalence results between testing procedures.

1.5 Appendix

1.5.1 Proof of lemma 1

For $\gamma > 0$, we have:

$$\phi'(m) = \frac{m^\gamma}{\gamma} \Longrightarrow (\phi')^{-1}(z) = [\gamma z]^{1/\gamma}$$

Moreover:

$$\begin{aligned} f(u) &= \phi(u) - \phi'(1)(u-1) \\ &\Longrightarrow f'(u) = \phi'(u) - \phi'(1) \end{aligned}$$

so that:

$$\begin{aligned} u &= (f')^{-1}(z) \Leftrightarrow z = \phi'(u) - \phi'(1) \\ &\Longrightarrow (f')^{-1}(z) = (\phi')^{-1}[z + \phi'(1)] = [\gamma z + \gamma\phi'(1)]^{1/\gamma} = [\gamma z + 1]^{1/\gamma} \end{aligned}$$

From Csiszár (1995):

$$\begin{aligned} f^*(v) &= \int_0^v (f')^{-1}(z)dz = \int_0^v [\gamma z + 1]^{1/\gamma} dz \\ &= \frac{1}{\gamma+1}\left\{[\gamma z + 1]^{\frac{\gamma+1}{\gamma}}\right\}_0^v \end{aligned}$$

Thus the conditions of Assumption A3 can be written for every $\alpha > 0$ and all $j = 1, ..., H$:

$$\begin{aligned} E\left\{f^*\left[\alpha Y_{j,\theta}^+\right]\right\} &= \frac{1}{\gamma+1}\left\{E\left[(\gamma\alpha Y_{j,\theta}^+ + 1)^{\frac{\gamma+1}{\gamma}}\right] - 1\right\} < \infty \\ E\left\{f^*\left[\alpha Y_{j,\theta}^-\right]\right\} &= \frac{1}{\gamma+1}\left\{E\left[(\gamma\alpha Y_{j,\theta}^- + 1)^{\frac{\gamma+1}{\gamma}}\right] - 1\right\} < \infty \end{aligned}$$

that is for all $j = 1, ..., H$:

$$\begin{aligned} E\left[(Y_{j,\theta}^{+})^{\frac{\gamma+1}{\gamma}}\right] &< \infty \\ E\left[(Y_{j,\theta}^{-})^{\frac{\gamma+1}{\gamma}}\right] &< \infty \end{aligned}$$

Since for all $p > 1$, $E[(X+1)^p] < \infty$ if and only if $E[X^p] < \infty$ since the $\mathbb{L}^p$ spaces are Banach. I.e since $1, -1 \in \mathbb{L}^p$, if $X + 1 \in \mathbb{L}^p$ then $X = (X+1) - 1 \in \mathbb{L}^p$ and vice versa. The above is true if and only if :

$$\begin{aligned} E[(|Y_{j,\theta}|)^{\frac{\gamma+1}{\gamma}}] =& E[\mathbf{1}_{y>0}(Y_{j,\theta}^{+})^{\frac{\gamma+1}{\gamma}}] + E[\mathbf{1}_{y<0}(Y_{j,\theta}^{-})^{\frac{\gamma+1}{\gamma}}] \\ =& E\left[(Y_{j,\theta}^{+})^{\frac{\gamma+1}{\gamma}}\right] + E\left[(Y_{j,\theta}^{-})^{\frac{\gamma+1}{\gamma}}\right] < \infty \end{aligned}$$

Where the second line follows since the functions are zero outside the indicators. This condition is equivalent to integrability of $|Y_{j,\theta}|^{\frac{\gamma+1}{\gamma}}$ which must be met for all of the individual functions a_j $j = 1, ..., H$.

For $\gamma = 0$ we have $\phi'(m) = ln(m) + 1$ so that:

$$f'(u) = \phi'(u) - \phi'(1) = ln(u) + 1 - ln(1) - 1 = ln(u)$$

This then implies $(f')^{-1}(z) = e^u$ so that:

$$\begin{aligned} f^*(v) &= \int_0^v (f')^{-1}(z)dz = \int_0^v e^z dz \\ &= \{e^z\}_0^v \\ &= e^v - 1 \end{aligned}$$

Applying this to the conditions of Assumption A3 we get:

$$
\begin{aligned}
E\left\{f^*\left[\alpha Y_{j,\theta}^{+}\right]\right\} &= E\left[e^{\alpha Y_{j,\theta}^{+}}\right] - 1 < \infty \\
E\left\{f^*\left[\alpha Y_{j,\theta}^{+}\right]\right\} &= E\left[e^{\alpha Y_{j,\theta}^{-}}\right] - 1 < \infty
\end{aligned}
$$

Which is true if and only if $E[e^{\alpha|Y_{j,\theta}|}] < \infty$ using the same reasoning as for $\gamma > 0$. This condition being true for all $\alpha > 0$ is then equivalent to:

$$E[e^{\beta Y_{j,\theta}}] < \infty \quad \forall \beta \in R$$

This must again hold true for all a_j $j = 1, ..., H$. Next notice that by the Cauchy-Schwarz inequality, where t_j denotes component j of the vector t in lemma 1:

$$E[exp(t'Y_\theta)] \leq \prod_{j=1}^{H} E[exp(2t_j Y_{j,\theta})]^{\frac{1}{2}}$$

So if the laplace transform of each component $Y_{j,\theta}$ is finite for all $\beta_j \in \mathbb{R}$, so must the laplace transform of the vector Y_θ for any $t \in \mathbb{R}^H$. If the laplace transform is finite, so must each individual component $Y_{j,\theta}$ have a finite laplace transform since we may use $t_k = 0 \; \forall k \neq j$. QED

1.5.2 Proof of lemma 2

Since:

$$E\left[\phi_\gamma(M^{(j)})\right] \geq 0, \forall j = 1, 2, ..$$

we only need to show that:

$$\limsup_{j=\infty} E\left[\phi_\gamma(M^{(j)})\right] \leq 0$$

that is:

$$\liminf_{j=\infty} E\left[(M^{(j)})^{\gamma+1}\right] \geq 1$$

We have by Jensen inequality for the concave function log :

$$\log\left\{E\left[(M^{(j)})^{\gamma+1}\right]\right\} \geq E\left\{(\gamma+1)\log\left\{M^{(j)}\right\}\right\}$$

Hence:

$$\lim_{j=\infty} E\left[\log(M^{(j)})\right] = 0 \Longrightarrow \liminf_{j=\infty} \log\left\{E\left[(M^{(j)})^{\gamma+1}\right]\right\} \geq 0$$

and by taking exponential:

$$\liminf_{j=\infty} \left\{E\left[(M^{(j)})^{\gamma+1}\right]\right\} \geq 1$$

QED

1.5.3 Proof of theorem 1

Case 1) $\gamma < -1$

As in the proof of lemma 2, we only need to show that:

$$\limsup_{j=\infty} E\left[\phi_\gamma(M^{(j)})\right] \leq 0$$

which, in the case of $\gamma < -1$, means:

$$\limsup_{j=\infty} E\left[(M^{(j)})^{\gamma+1}\right] \leq 1$$

By definition:

$$M_d^{(j)} \leq M^{(j)} \leq M_u^{(j)}$$

where:

$$\begin{aligned} M_d^{(j)} &= 1-\pi_j \\ M_u^{(j)} &= 1-\pi_j+\frac{\pi_j}{l_j(\theta)}H_j(\theta) \\ H_j(\theta) &= \sup\left\{\frac{1}{h_\theta(y)}; y\in G_\theta^{\varepsilon_j}(\alpha_j)\right\} \end{aligned}$$

Note that $H_j(\theta)$ is finite since h_θ is continuous and strictly positive on the compact set $G_\theta^{\varepsilon_j}(\alpha_j)$.

Let us denote:

$$M^{(j)}=\beta_j M_d^{(j)}+(1-\beta_j)\,M_u^{(j)}, \beta_j\in[0,1]$$

By convexity of the function $g_\gamma(x)=x^{\gamma+1}$, we have:

$$\begin{aligned} g_\gamma\left(M^{(j)}\right) &\leq \beta_j g_\gamma\left(M_d^{(j)}\right)+(1-\beta_j)\,g_\gamma\left(M_u^{(j)}\right) \\ &= g_\gamma\left(M_d^{(j)}\right)+(1-\beta_j)\left[g_\gamma\left(M_u^{(j)}\right)-g_\gamma\left(M_d^{(j)}\right)\right] \end{aligned}$$

Note that:

$$\lim_{j=\infty} g_\gamma\left(M_d^{(j)}\right)=\lim_{j=\infty}\left[1-\pi_j\right]^{\gamma+1}=1$$

since the sequence π_j converges towards zero. Therefore, we only need to show that:

$$\lim_{j=\infty}\sup\left[g_\gamma\left(M_u^{(j)}\right)-g_\gamma\left(M_d^{(j)}\right)\right]\leq 0$$

Note that by definition:

$$H_j(\theta)\geq\frac{1}{h_\theta\left[-\alpha_j E\left(Y_\theta\right)\right]}$$

and:

$$\lim_{j=\infty} h_\theta\left[-\alpha_j E\left(Y_\theta\right)\right]=0$$

since $E\left(Y_\theta\right)\neq 0$, the sequence λ_j converges to infinity and the non-negative integrable

function h_θ must go to zero at infinity. Thus, since $\gamma + 1 < 0$:

$$\lim_{j=\infty} H_j(\theta) = +\infty \Longrightarrow \lim_{j=\infty} M_u^{(j)} = +\infty \Longrightarrow \lim_{j=\infty} g_\gamma\left(M_u^{(j)}\right) = 0$$

Hence:

$$g_\gamma\left(M_d^{(j)}\right) \geq 0 \Longrightarrow \limsup_{j=\infty}\left[g_\gamma\left(M_u^{(j)}\right) - g_\gamma\left(M_d^{(j)}\right)\right] \leq 0$$

QED

Case 2) $\gamma = -1$

We need to show:

$$\limsup_{j=\infty} E\left[-\log(M^{(j)})\right] \leq 0$$

that is:

$$\liminf_{j=\infty} E\left[\log(M^{(j)})\right] \geq 0$$

Let us denote:

$$\pi_j^0(\theta) = \Pr\left[Y_\theta \in G_\theta^{\varepsilon_j}(\alpha_j)\right]$$

We will use several times the fact that, by choosing $\varepsilon_j \leq \bar{\varepsilon}$, we have by virtue of assumption A4(ii):

$$\pi_j^0(\theta) \leq BVol\left[G_\theta^{\varepsilon_j}(\alpha_j)\right]$$

where $Vol\left[G_\theta^{\varepsilon_j}(\lambda_j)\right] = l_j$, the volume of the ball $G_\theta^{\varepsilon_j}(\alpha_j)$, is a function $V(\theta, \varepsilon_j)$, independent of α_j and such that:

$$\lim_{\varepsilon \to 0} V(\theta, \varepsilon) = 0 \tag{1.14}$$

Moreover:

$$\sup_{j \geq 1} l_j(\theta) = \bar{l}(\theta) \leq V(\theta, \bar{\varepsilon})$$

By definition:

$$E\left[\log(M^{(j)})\right] = \left(1-\pi_j^0(\theta)\right)\log\left[1-\pi_j\right] + \int_{G_\theta^{\varepsilon_j}(\alpha_j)} \log\left\{1-\pi_j+\frac{\pi_j}{l_j(\theta)}\frac{1}{h_\theta(y)}\right\} h_\theta(y)dy$$

Since π_j converges to zero, we only have to show that:

$$\liminf_{j=\infty} \int_{G_\theta^{\varepsilon_j}(\alpha_j)} \log\left\{1-\pi_j+\frac{\pi_j}{l_j(\theta)}\frac{1}{h_\theta(y)}\right\} h_\theta(y)dy \geq 0$$

However:

$$\log\left\{1-\pi_j+\frac{\pi_j}{l_j(\theta)}\frac{1}{h_\theta(y)}\right\} \geq \log\left\{\frac{\pi_j}{l_j(\theta)}\frac{1}{h_\theta(y)}\right\} \geq \log\left\{\frac{\pi_j^0}{l_j(\theta)}\frac{1}{h_\theta(y)}\right\}$$

where the second inequality is warranted since, by virtue of (1.14), we get $\pi_j^0 \leq \pi_j$ by choosing ε_j small enough. Thus, we only need to show that:

$$\liminf_{j=\infty} \int_{G_\theta^{\varepsilon_j}(\alpha_j)} \log\left\{\frac{\pi_j^0}{l_j(\theta)}\frac{1}{h_\theta(y)}\right\} h_\theta(y)dy \geq 0$$

Note that, by (1.14), when choosing ε_j going to zero when $j \to \infty$, we also have $l_j(\theta)$ going to zero, so that for j large enough:

$$\begin{aligned}\int_{G_\theta^{\varepsilon_j}(\alpha_j)} \log\left\{\frac{1}{l_j(\theta)}\right\} h_\theta(y)dy &= \int_{G_\theta^{\varepsilon_j}(\alpha_j)} \left|\log\left[l_j(\theta)\right]\right| h_\theta(y)dy \leq M\log\left[l_j(\theta)\right] l_j(\theta) \\ &\Longrightarrow \lim_{j=\infty} \int_{G_\theta^{\varepsilon_j}(\alpha_j)} \log\left\{\frac{1}{l_j(\theta)}\right\} h_\theta(y)dy = 0\end{aligned}$$

Hence, we only have to show that:

$$\liminf_{j=\infty} \int_{G_\theta^{\varepsilon_j}(\alpha_j)} \log\left\{\frac{\pi_j^0}{h_\theta(y)}\right\} h_\theta(y)dy \geq 0$$

However:

$$\liminf_{j=\infty} \int_{G_\theta^{\varepsilon_j}(\alpha_j)} \log\left\{\pi_j^0\right\} h_\theta(y)dy = \liminf_{j=\infty} \pi_j^0 \log(\pi_j^0) = 0$$

since $\pi_j^0 \leq \pi_j$ goes to zero like π_j. Hence we only have to show that:

$$\limsup_{j=\infty} \int_{G_\theta^{\varepsilon_j}(\alpha_j)} \log\left\{h_\theta(y)\right\} h_\theta(y)dy \leq 0$$

This inequality is directly implied by assumption A4 (iii), since for j large enough:

$$\alpha_j = \frac{1-\pi_j}{\pi_j} > \bar{\alpha} \Longrightarrow \log\left\{h_\theta(y)\right\} \leq 0, \forall y \in G_\theta^{\varepsilon_j}(\alpha_j)$$

QED

CHAPTER 2

BOUNDED TILTING ESTIMATION

2.1 Introduction

We are concerned with a moment condition model using a H dimensional vector valued non-linear moment function $g(X, \theta)$ where X is a random vector and $\theta \in \Theta$ a finite dimensional vector of parameters of interest. The moment condition pins down a relationship between the true parameter value and the moments of the distribution of the random vector X through the following equation:

$$\mathbb{E}[g(X, \theta_0)] = 0 \tag{2.1}$$

We say that the moment condition model is well specified (resp. misspecified) if such a θ_0 satisfying (2.1) exists (resp. if such a θ_0 does not exist). The methodology of estimating θ_0 in a well specified model has seen change throughout the last couple of decades. Hansen (1982) introduced the desirable GMM estimator, which remained standard practice until it was pointed out that the finite sample properties are not optimal. New estimators were proposed which take advantage of the moment condition (2.1) to improve the finite sample performance. These include Empirical Likelihood (EL), Exponential Tilting (ET), and Continuous Updating GMM (CUE). These estimators are members of a subclass of the Minimum Divergence Estimators or Minimum Discrepancy Estimators (MD) (well explained by Corcoran (1998)), which this chapter will now set its focus on.

The MD estimators are first order equivalent to two-step efficient GMM under regularity conditions, implying that they reach the semi-parametric efficiency bound. They are also one-step estimators so that reliance on an arbitrary first step estimator is unnec-

essary and further they possess great computational properties since they can be written as the solution to a saddle point problem. A persistent interest of the literature in recent years has then been to determine which estimators in the class are most desirable. A particular focus has been set on the Empirical Cressie-Read Estimators (ECR), of which the aforementioned EL, ET, and CUE are members. Since these estimators all share the same desirable properties mentioned above, the exploration of how to further narrow them down has been funneled into two directions.

First there is the question of higher order efficiency. This was first explored by Newey and Smith (2004) who showed that bias-corrected EL is second order efficient in the class of Generalized Empirical Likelihood (GEL) estimators (a large subclass of MD estimators). Later Ragusa (2011) generalized the idea and showed that bias-corrected EL is second order efficient in the entire class of MD estimators. This analysis suggests that EL is the most desirable estimator from this perspective. The second point of comparison is how the estimators stand up to model misspecification. Imbens et al. (1995) argued that EL would suffer in a misspecified model due to an asymptote in its influence function, something which was later formalized by Schennach (2007) who showed that EL is not $\sqrt{n}$ consistent in a misspecified model when the moment function g is unbounded. Schennach also showed that ET is robust to misspecification, suggesting that it is the most desirable estimator from this perspective. Schennach also suggested that by using a combination estimator Exponentially Tilted Empirical Likelihood (ETEL), one can retain the higher order properties of EL while also being robust to misspecification like ET.

The problem is that to gain robustness to misspecification, heavy assumptions must be made on the DGP X and on the moment function g to ensure that the asymptotic reweighting problem which the MD class uses has a solution. Without such a solution the concept of robustness is ill-defined and can hence not be explored. An exploration of sufficient conditions which guarantee existence of such a solution for the ECR class is given by Renault and Wahlstrom in Chapter 1 of this Dissertation. We showed that

depending on the choice of divergence function different assumptions must be made on the existence of moments of the transformed random variable $Y_\theta = g(X, \theta)$ for all fixed θ. ET and ETEL both require a bounded moment generating function for Y_θ. The purpose of this chapter is to produce a class of estimators in the MD class which gets around this issue and is robust without assumptions on the DGP, while also retaining the higher order efficiency of the EL estimator.

Since the ECR subclass is well explored and understood and none of the members have an easy way to get around the issue, to do this we move beyond this subclass into the full class of MD estimators. A simple way around the misspecification problem turns out to be to choose estimators that have a bounded tilting function. This ensures that robustness is possible with the only assumption being that the random variable $Y_\theta = g(X, \theta)$ is absolutely continuous for all $\theta \in \Theta$, which is also necessary to assume for the other estimators. Within this class, there are natural suggestions of tilting functions which satisfy all the necessary requirements - combinations of scaled CDFs with full support on $\mathbb{R}$. By choosing the CDFs in the appropriate manner one can also attain the same higher order efficiency as EL just like ETEL, but without any assumptions on the DGP other than the usual regularity conditions. In other words, we present a large subclass of MD estimators which are heretofor unexplored and which have the optimality properties which both strands of literature have been searching for.

The rest of this chapter is organized as follows. Section 2 reviews the Minimum Divergence Estimators, their first and higher order efficiency, and results about misspecification. Section 3 defines the bounded tilting estimators, proves their higher order efficiency and their behavior under misspecification. Section 4 gives an example of an easily implementable tilting function and shows its performance in simulations and Section 5 concludes.

2.2 Review of Key Concepts in Minimum Divergence Estimation

2.2.1 The Estimator

To exploit the moment condition (2.1) in the sample, the MD estimators fix $\theta \in \Theta$ and minimally reweight the data points $\{x_i\}_{i=1}^n$ using a weight vector $w = \{w_i\}_{i=1}^n$ so that a sample counterpart to this moment condition holds true while the weights sum to 1. For each θ we then ask "how much" reweighting is necessary to achieve this and then the θ which requires the least amount of reweighting is our estimator $\hat{\theta}_q$, where the amount is defined using a divergence function q.

Formally, the Minimum Divergence Estimators $\hat{\theta}_q$ are defined using a strictly convex twice differentiable divergence function q in the following optimization problem:

$$\begin{aligned} \hat{\theta}_q = argmin_{\theta\in\Theta} min_{\{w_i\}} \frac{1}{n}\sum_{i=1}^{n} q(nw_i) \qquad (2.2) \\ s.t. \quad \sum_{i=1}^{n} w_i g(x_i,\theta) = 0 \\ \sum_{i=1}^{n} w_i = 1 \end{aligned}$$

Where $q(1) = \dot{q}(1) = 0$ and $\frac{d^2}{dx^2}q(1) = 1$. For clarity, when we reference a divergence function in the rest of this chapter we will implicitly assume it is strictly convex and twice continuously differentiable. The ECR family we referred to in the introduction is the family of divergence functions q_γ indexed by γ and we write it here for reference:

$$q_\gamma(x) = \begin{cases} \frac{x^{\gamma+1}-1}{\gamma(\gamma+1)}, & \forall \gamma \neq 0, -1 \\ -\log(x), & \gamma = -1 \\ x\log(x), & \gamma = 0 \end{cases}$$

Denoting by λ the Lagrange multiplier for the first constraint and μ the Lagrange multiplier for the second constraint we may write the Lagrangian as:

$$L(\theta, w, \lambda, \mu) = \frac{1}{n}\sum_{i=1}^{n} q(nw_i) - \lambda' \sum_{i=1}^{n} w_i g(x_i, \theta) - \mu(\sum_{i=1}^{n} w_i - 1) \tag{2.3}$$

An interior solution to (2.2) must set all the partial derivatives of L to 0. Denoting by $G(x,\theta) = \frac{\partial g(x,\theta)}{\partial \theta'}$, and $\dot{q}(x) = \frac{dq(x)}{dx}$ we get from the partial derivatives of θ and w_i respectively:

$$\sum_{i=1}^{n} w_i \lambda' G(x_i, \theta) = 0$$

and

$$\dot{q}(nw_i) - \lambda' g(x_i, \theta) - \mu = 0 \quad \forall i = 1, ..., n$$

The partial derivative with respect to w_i we may then use to yield a closed form for w_i in terms of the Lagrange multipliers and the moment function g:

$$w_i = \frac{1}{n}\dot{q}^{-1}(\mu + \lambda' g(x_i, \theta))$$

The mapping $\dot{q}^{-1}$ is known as the tilting function, which determines the optimal weights from the value of the Lagrange multipliers and the value of the moment function. We may use this to rewrite the constraints into

$$\frac{1}{n}\sum_{i=1}^{n}\dot{q}^{-1}(\mu+\lambda' g(x_i,\theta))g(x_i,\theta)=0, \qquad \frac{1}{n}\sum_{i=1}^{n}\dot{q}^{-1}(\mu+\lambda' g(x_i,\theta))=1$$

This together with the partial derivative with respect to θ gives us a set of estimating equations:

$$\sum_{i=1}^{n}\rho(x_i,\hat{\theta}_q,\hat{\lambda}_q,\hat{\mu}_q)=0$$

Where:

$$\rho(x,\theta,\lambda,\mu)=\begin{bmatrix}\dot{q}^{-1}(\mu+\lambda' g(x,\theta))\lambda' G(x,\theta)\\ \dot{q}^{-1}(\mu+\lambda' g(x,\theta))g(x,\theta)\\ \dot{q}^{-1}(\mu+\lambda' g(x,\theta))-\frac{1}{n}\end{bmatrix} \tag{2.4}$$

In addition to the estimating equation formulation above, a big advantage of the MD estimators is that they allow for a saddle point problem formulation. As is shown in Ragusa (2011) on page 7 equation (4) the dual problem to (2.2) has the following formulation:

$$\hat{\theta}_q=argmax_{\theta\in\Theta}min_{(\mu,\lambda)\in\Lambda_n(\theta)}\frac{1}{n}\sum_{i=1}^{n}q^*(\mu+\lambda' g(x_i,\theta))-\mu \tag{2.5}$$

With

$$\Lambda_n(\theta)=\{(\mu,\lambda'):\mu+\lambda' g(x_i,\theta)\in Dom(q^*),\forall i=1,...,n\}$$

And where q^* is the convex conjugate of q:

$$q^*(v):=sup_{u\in dom(q)}[uv-q(u)]$$

The function q^* must then satisfy $q^*(0)=0$ and $\frac{d}{dx}q^*(0)=\frac{d^2}{dx^2}q^*(0)=1$ and just like q be twice continuously differentiable and strictly convex. The inner optimization problem

over (μ, λ) is different from that of its primal in that it only requires $dim(g(.,.)) + 1$ parameters to be estimated, unlike its primal counterpart which requires n parameters (one weight for each data point). This makes the estimation procedure much faster and easier to execute. In our simulations later on in this chapter we use the dual formulation to produce our results.

Remark: Note the presence of the $-\mu$ at the end of the optimization problem in (2.5). This is there to generate the equivalence between the FOC of (2.5) and the FOC of (2.2). In the latter we have:

$$\frac{1}{n}\sum_{i=1}^{n}\dot{q}^{-1}(\mu + \lambda' g(x_i, \theta)) = 1$$

Which corresponds to the partial derivative of (2.5) with respect to μ.

2.2.2 First and Higher Order Asymptotics

We now briefly summarize the results from Ragusa Ragusa (2011) on the asymptotics of the MD estimators that are relevant for our chapter. We refer our readers to his chapter for more details on the topics. All the results in this section are valid for the ECR estimators as well since they are also MD estimators. We first recite the assumptions for the first order asymptotics:

Assumption 1. *(a)* $\theta_0 \in \Theta$ *is the unique solution to* $\mathbb{E}[g(X, \theta)] = 0$*; (b)* Θ *is compact; (c)* $g(., \theta)$ *is continuous in* θ *at all* $\theta \in \Theta$ *w.p.1; (d)* $\mathbb{E}[sup_\theta |g(X, \theta)||^2] < \infty$*; (e)* $\Omega = \mathbb{E}[g(X, \theta_0)g(X, \theta_0)']$ *is non-singular*

Assumption 2. *(a)* $\theta \in int(\Theta)$*; (b)* $g(x, \theta)$ *is continuously differentiable in a neighborhood* $\mathcal{N}$ *of* θ_0*; (c)* $\mathbb{E}[sup_{\theta \in \mathcal{N}} ||G(X, \theta)||] < \infty$*; (d)* $Rank(G) = dim(\theta)$*,* $G = \mathbb{E}[G(X, \theta_0)]$

Under these assumptions Ragusa shows that the MD estimators have the same first order properties as the GEL estimators (which use a generalized homogeneous tilting function), and those are equivalent to two-stage GMM and are hence efficient:

Theorem 1. *(Ragusa Ragusa (2011) Theorems 5-6 p.12)*

1) Under Assumption 1 we have:

$\hat{\theta}_q \to_p \theta_0$, $\hat{\mu}_q = o_p(n^{-\frac{1}{2}})$, *and* $\hat{\lambda}_q = O_p(n^{-\frac{1}{2}})$

2) Under Assumptions 1 and 2 we have:

$$\sqrt{n}\begin{pmatrix} \hat{\lambda}_q \\ \hat{\theta}_q - \theta_0 \end{pmatrix} \to_d \mathcal{N}\left(0, \begin{pmatrix} P & 0 \\ 0 & \Sigma \end{pmatrix}\right)$$

where $\Sigma = (G'\Omega^{-1}G)^{-1}$ *and* $P = \Omega^{-1}(I_M - G\Sigma G'\Omega^{-1})$ *with* $M = dim(g(.,.))$

Ragusa also conducts a thorough investigation of the higher order properties of the MD class in his section 5, just as Newey and Smith (2004) who base their work on the discussion of Rothenberg (1984). The starting point is the $O_p(n^{-2})$ expansion of an estimator $\hat{\theta}$:

$$(\hat{\theta} - \theta_0) = \frac{i_n}{\sqrt{n}} + \frac{b_n}{n} + \frac{c_n}{\sqrt{n}n} + \frac{r_n}{n^2}$$

Where i_n, b_n, c_n, and r_n are $O_p(1)$. This allows Ragusa to define first the higher order bias (of order $O(n^{-1})$) of $\hat{\theta}$ as:

$$Bias_1(\hat{\theta}) = \frac{\mathbb{E}[b_n]}{n}$$

Secondly Ragusa defines the $O(n^{-2})$ Mean Squared Error, denoted by $MSE_2(\hat{\theta})$, of $\hat{\theta}$ as:

$$MSE_2(\hat{\theta}) = \frac{\mathbb{E}[i_n i_n']}{n} + \frac{\Xi}{n}$$

where

$$\Xi := \mathbb{E}[\frac{b_n b_n'}{n}] + \mathbb{E}[(\frac{b_n}{\sqrt{n}} + \frac{c_n}{n})i_n')] + \mathbb{E}[i_n(\frac{b_n}{\sqrt{n}} + \frac{c_n}{n})']$$

The higher order efficiency of the estimators is in regards to MSE_2, but to exclude unreasonable estimators from consideration Ragusa first limits his discussion to Bias Corrected estimators $\hat{\theta}^c = \hat{\theta} - \frac{\widehat{\mathbb{E}[b_n]}}{n}$. He then defines higher order efficiency, just like Newey and Smith (2004), as:

Definition 1. (Ragusa (2011) Definition 5 p.17) The bias corrected estimator $\hat{\theta}^c$ is second order efficient if for any other bias corrected estimator $\overline{\theta}^c$ there exists a positive definite matrix Π such that $MSE_2(\hat{\theta}^c) - MSE_2(\overline{\theta}^c) = \Pi + o(n^{-2})$.

To prove his higher order efficiency result he needs one more set of assumptions:

Assumption 3. *There is $b(x)$ with $\mathbb{E}[||b(x)||^6] < \infty$ such that for $0 \leq j \leq 4$ and all x, $\frac{\partial^j g(x,\theta)}{\partial \theta^j}$ exists on a neighborhood $\mathcal{N}$ of θ_0, $sup_{\theta \in \mathcal{N}} || \frac{\partial^j g(x,\theta)}{\partial \theta^j} || \leq b(x)$, and for each $\theta \in \mathcal{N}$ $|| \frac{\partial^4 g(x,\theta)}{\partial \theta^4} - \frac{\partial^4 g(x,\theta_0)}{\partial \theta^4} || \leq b(x)||\theta - \theta_0||$.*

Also, q^* is four times continuously differentiable with Lipschitz fourth derivative in a neighborhood of zero.

Under these assumptions Ragusa shows that $q_3^* = \frac{\partial^3}{\partial x^3} q^*(0)$ is the only thing which matters for second order efficiency. In particular, any bias corrected estimator which has the same q_3^* will also have the same higher second order MSE. This together with the fact that Newey (2004) have shown that bias corrected EL is second order efficient in the sense of our above definition (and $q_3^* = 2$ for EL) yields:

Theorem 2. *(Ragusa (2011) Corollary 1 p.21) If assumptions 1, 2, and 3 hold then all bias corrected MD estimators with $q_3^* = 2$ are second order efficient in the sense of the above definition.*

The main take aways from this section are then that all MD estimators are just as good as GMM in the first order, and if $q_3^* = 2$ they are also second order efficient. Note that while the normalizations $q^*(0) = 0$ and $\frac{d}{dx} q^*(0) = \frac{d^2}{dx^2} q^*(0) = 1$ are inconsequential in defining the MD estimators since you can always rescale them to satisfy these conditions,

these normalizations are necessary in the above theorem since otherwise the optimality condition of $q_3^* = 2$ will change.

Remark: Note that EL is the only member of the class of ECR estimators which is second order efficient.

Remark: Newey and Smith (2004) require $q_4^* = 6$ for this result but Ragusa shows that this is not necessary. Similarly to the situation for the ETEL estimator of Schennach (2007) the difference in the third order terms between the estimators is uncorrelated with the first order term so that the higher order variances are the same.

2.2.3 Misspecification, Pseudo-True Values, and Robustness

Let us now formalize what we mean with robustness to misspecification. Should there not exist any $\theta \in \Theta$ for which $\mathbb{E}[g(X,\theta)] = 0$, the model is misspecified. We may then ask which θ is the "closest" to satisfying the moment constraints in the population, where closest refers to the least divergence according to some divergence function q. This θ_q^* is known as the pseudo-true value and is the solution to the population optimization problem:

$$\begin{aligned} \theta_q^* = argmin_{\theta\in\Theta} min_M \mathbb{E}[q(M)] \\ s.t. \quad \mathbb{E}[Mg(X,\theta)] = 0 \\ \mathbb{E}[M] = 1 \end{aligned} \tag{2.6}$$

Note here that the M are functions of $g(X,\theta)$, and in particular they are change of measure random variables. Also note that the pseudo-true value also depends on the choice of divergence function q. We say that an estimator is robust to misspecification if $|\hat{\theta}_q - \theta_q^*| = O_p(\frac{1}{\sqrt{n}})$. However, as pointed out by Renault and Wahlstrom in Chapter 1 of this dissertation, the existence of such a θ_q^* is far from granted. In order for robustness

to be possible such a θ_q^* must exist so we will explore what is necessary for θ_q^* to exist. Following the logic in Chapter 1 such a θ_q^* existing necessitates the existence of a solution to the reweighting problem:

$$\begin{aligned} M_\theta &= argmin_M \mathbb{E}[q(M)] \\ s.t. \quad \mathbb{E}[Mg(X,\theta)] &= 0 \\ \mathbb{E}[M] &= 1 \end{aligned} \tag{2.7}$$

for all $\theta \in \Theta$. Renault and Wahlstrom show in Chapter 1 that the existence of such an M_θ necessitates assumptions on the divergence function q and on the random variable X for a fixed θ using the work of Csiszár (1995). These conditions are then translated to the Cressie-Read family of q.

As is shown in Chapter 1, Csiszár's work is applicable to the entire class of MD estimators. For existence of such a change of measure random variable M_θ we need conditions both on q and on the random variable $Y_\theta = g(X,\theta)$ for every fixed θ. The conditions necessary depend on what assumptions one is willing to make on Y_θ, since there is a trade-off in the assumptions made on Y_θ and how stringent one must be with q. If the random variable Y_θ is bounded, which is the strongest assumption possible, the only conditions necessary to apply Csiszár's results are:

Condition 1. *The random variable Y_θ is absolutely continuous with respect to some σ-finite measure for every fixed θ*

Condition 2. $q(1) = \dot{q}(1) = 0$

As is explained in Renault and Wahlstrom (2020), condition 1 is there to make the optimization problem at hand (2.7) fit into Csiszár's work. Condition 2 simply ensures that the function q creates a distance when integrated over. This is exactly the condition

which we impose on the optimization problem (2.2) that define the MD estimators. Note that there is no need to place any condition on the second derivative of q here for Csiszár's results to hold. The imposition of these conditions are just to simplify the analysis since any function that is twice continuously differentiable and strictly convex can be rescaled to satisfy these conditions arbitrarily. The reason we maintained the second derivative normalization in the previous subsections was to get the optimality condition for Theorem 2 but the rescaling plays no role here. Under these conditions Csiszár shows:

Theorem 3. *(Csiszár (1995) Theorem 3 ii) p.177)*

Under conditions 1 and 2, M_θ exists for every θ if Y_θ is bounded in every direction for every θ.

To move beyond the boundedness assumption we need to add two more conditions, one on q and one on the random variable Y_θ. Denote by $Y_{j,\theta}$ the $j-$th component of the vector Y_θ. Also, let $dom(\dot{q}) = (0, a)$ with $a \in (1, \infty]$. These conditions are then:

Condition 3. $lim_{x \to a} \dot{q}(x) = \infty$

Condition 4. $\mathbb{E}[q^*(\alpha|Y_{j,\theta}|)] < \infty$ *for all* $\alpha > 0$ *and* $j = 1, .., H$ *for every* $\theta \in \Theta$

where q^* is once again the convex conjugate of q:

$$q^*(v) = sup_{u \in dom(q)}[uv - q(u)]$$

Under these conditions Csiszár also proves the existence of M_θ.

Theorem 4. *(Csiszár (1995) Theorem 3 iii) p.177)*

Under conditions 1,2,3, and 4, M_θ exists for every $\theta \in \Theta$.

We now explain the conditions that are used. Condition 3 can be understood by analogy to the first order conditions which come from (2.7), explained in Chapter 1. The solution M_θ must satisfy the following first order condition almost surely (see Chapter 1 page 11 equation 1.10):

$$\dot{q}(M_\theta) = \mu^* + \lambda^* Y_\theta$$

where μ^* and λ^* are the values of the Lagrange multipliers that are pinned down from setting the partial derivatives of the Lagrangian to zero under mild differentiability conditions.

If Y_θ is unbounded then the RHS of this equation is also unbounded and can diverge to $+\infty$. Since $\dot{q}$ is strictly increasing (since q is strictly convex), if we want the LHS to also diverge to $+\infty$ we must let M_θ move to the upper bound of the domain of $\dot{q}$ and we need $\dot{q}$ to diverge to $+\infty$ there.

Remark: Notice that it is tempting to use the same analogous explanation to suggest that one should also impose the condition $lim_{x\to 0}\dot{q}(x) = -\infty$, as Ragusa (2011) suggests. It turns out however that this is not necessary for the existence of a solution to the problem but rather to ensure that the solution M_θ is strictly positive (almost surely). To understand why this is the case consider what happens when Y_θ diverges to $-\infty$ for a given θ. Since $\dot{q}$ is strictly increasing, this must be coupled with M_θ moving towards zero. However, when it reaches zero the constraint associated with λ^* ($\mathbb{E}[MY_\theta)] = 0$) is automatically fulfilled so the Lagrange multiplier becomes zero. This then means that as long as $lim_{x\to 0}\dot{q}(x) = \mu^*$ the FOC holds.

Condition 4 comes from a clever derivation by Cziszar using Orlicz spaces, but is beyond the scope of this chapter. We provide instead some intuition behind it as follows. For the minimizer M_θ in (2.7) to exist we must guarantee that it satisfies $\mathbb{E}[M_\theta Y_\theta] = 0$, while the only thing we know is that $\mathbb{E}[q(M)] < \infty$ for all M under consideration in the optimization problem. We then need to ensure that any sequence of M_j which converges to M_θ satisfies $\mathbb{E}[M_j Y_\theta] \to \mathbb{E}[M_\theta Y_\theta]$, which implies what we need. If the sequence M_j converges in L^1 this is guaranteed, and condition 4 gives exactly this convergence.

We finish this section with the fact that the existence theorems of the change of measure random variable M_θ coupled with a continuity arguement can then guarantee

existence of a pseudo-true value by an application of Berge's Maximum Theorem:

Corollary 1. Let $Z = \cup_\theta Z_\theta$ be the union of Z_θ, the spaces of absolutely continuous random variables with respect to the same σ-finite measure as Y_θ. Assume that conditions 1 and 2 hold for Y_θ bounded and 1,2,3, and 4 hold for Y_θ unbounded.
If the correspondence $C : \Theta \rightrightarrows Z$ with $C(\theta) = \{M \in Z : \mathbb{E}[MY_\theta] = 0 \; and \; \mathbb{E}[M] = 1\}$ is continuous and Θ is compact, θ_q^* exists.

Proof. Under the conditions we have that M_θ exists for every θ. The correspondence being continuous makes M_θ continuous by Berge's Maximum Theorem. The optimization problem which defines θ_q^* is then simply minimizing M_θ over Θ, where M_θ is continuous and Θ is compact - which guarantees the existence of θ_q^* by the Extreme Value Theorem. □

Remark: The assumptions necessary for the correspondence C to be continuous vary with Y_θ since the only thing which changes with θ is the condition $\mathbb{E}[MY_\theta] = 0$.

Remark: The existence of a pseudo-true value does not guarantee robustness, which we defined as $\sqrt{n}$ consistency towards the pseudo-true value by itself. For robustness of $\hat{\theta}_q$ one would need additional assumptions to guarantee the uniform convergence of the optimization problem (2.2) to (2.6). This is however immaterial for the discussion since the main problem at hand in the robustness question is whether there is something to converge towards in the first place.

2.3 Beyond ECR - Bounded Tilting

Now that we are equipped with the tools necessary, let us first explain the choice to depart from the ECR estimators. When applying Csiszár's conditions to the ECR estimators, Renault and Wahlstrom show in Chapter 1 that the only estimators in the class which can have a pseudo-true value when Y_θ is unbounded are the ones with $\gamma \geq 0$. Condition 4 in the ECR setting is equivalent to $\mathbb{E}[|Y_{j,\theta}|^{\frac{\gamma+1}{\gamma}}] < \infty$ for all j when $\gamma > 0$ and $\mathbb{E}[exp(tY_{j,\theta})] < \infty$

for all j and $t \in \mathbb{R}$ when $\gamma = 0$. When $\gamma < 0$, condition 3 does not hold and Renault and Wahlstrom show that no solution can exist to the optimization problem.

In other words, for the change of measure random variable M_θ to exist for the ECR class either boundedness or one of the above conditions must be assumed. In addition, the only member of the ECR class which is higher order efficient is EL (with $\gamma = -1$). The work of Renault and Wahlstrom then effectively proves that no estimator in the class can be both higher order efficient as well as be robust to misspecification with somewhat reasonable assumptions (as pointed out by Schennach (2007), a bounded moment function is not reasonable) at the same time. Our goal now will be to move beyond ECR to achieve this, as well as make the assumptions on the DGP as weak as possible.

We will restrict ourselves to Y_θ that satisfy condition 1 of Csiszár, and see how weak we can make any additional assumptions on Y_θ to guarantee existence of M_θ. As long as our q fits in the MD class, condition 2 is fulfilled so we must pick a q which gives us conditions 3 and 4. The condition which includes Y_θ is condition 4 and the only way that we can guarantee it without making any assumptions on the GDP is to ensure that q^* is bounded, which should not be surprising given the trade-off between assumptions on q and on Y_θ. Choosing a bounded q^* is facilitated by the following lemma:

Lemma 1. *For q^* twice continuously differentiable and strictly convex we have:*

$$q^*(v) = \int_0^v \dot{q}^{-1}(x)dx$$

For all $v \in Dom(q)$.

Proof. Notice that $q^*(v) = v\dot{q}^{-1}(v) - q(\dot{q}^{-1}(v))$, where $\dot{q}^{-1}$ exists because of the twice continuous differentiability and strict convexity, and taking its derivative yields:

$$\dot{q}^*(v) = \dot{q}^{-1}(v) + v\frac{\partial}{\partial v}\dot{q}^{-1}(v) - \frac{\partial}{\partial v}\dot{q}^{-1}(v)\dot{q}((\dot{q}^{-1}(v))$$

Cancelling the last two terms yields the equality:

$$\dot{q}^*(v) = \dot{q}^{-1}(v)$$

Then normalizing $q^*(0) = 0$ finishes the proof □

Lemma 1 shows that there is an intimate link between the convex conjugate q^* and the tilting function $\dot{q}^{-1}$. By imposing conditions on q so that we get a bounded tilting function we are then implicitly also imposing conditions on q^*. This brings us to the definition of our Bounded Tilting Estimators:

Definition 2. A Bounded Tilting Estimator is a MD estimator which uses a bounded tilting divergence function q_b which satisfies:

a) q_b is twice continuously differentiable and

b) $dom(\dot{q}_b) = (0, a)$ with $a \in (1, \infty)$

c) $lim_{x \to a} \dot{q}_b(x) = \infty$

The three conditions together ensure that a) the tilting function $\dot{q}_b^{-1}$ is well defined, b) is bounded, and c) allows the first order conditions to hold respectively. We then know that:

Lemma 2. *For a twice continuously differentiable strictly convex function q, let $Dom(q)$ be bounded. Then if $\dot{q}^{-1}$ is bounded, q^* is also bounded.*

Proof. Twice continuously differentiable means that $\dot{q}$ is continuous and strictly convex implies that $\dot{q}$ is strictly increasing which implies that $\dot{q}$ is invertible. Hence, $Range(\dot{q}^{-1}) = Dom(q)$ and if $Dom(q)$ is bounded then $Range(\dot{q}^{-1})$ is bounded.

By Lemma 1, q^* is the integral of $\dot{q}^{-1}$ and integrating a bounded function over a bounded domain yields a bounded integral. Hence, q^* is bounded. □

We can then simply apply Csiszár's theorems to yield:

Theorem 5. *When Y_θ is absolutely continuous with respect to a σ-finite measure for all θ, M_θ exists for all θ for all bounded tilting divergence functions q_b .*

Proof. The proof is simply verifying that the q_b satisfy Csiszár's conditions explained in the previous section, and applying theorem 4.

Condition 1 is assumed, condition 2 holds because the bounded tilting estimators are MD estimators, condition 3 is given from c) in the definition of q_b, and finally condition 4 follows from lemmas 1 and 2, since the expectation of a bounded function is also bounded. Hence by Theorem 4 M_θ exists. □

Since all Bounded Tilting Estimators are equally well-equipped to deal with misspecification, we will now narrow down the scope to the ones which are also higher order efficient in the sense of section 2.3. What we require to apply Theorem 2 are the necessary differentiability properties together with $\frac{\partial^3}{\partial x^3} q_b^*(x)|_0 = 2$. By lemma 1 the latter is implied by $\frac{\partial^2}{\partial x^2} \dot{q}_b^{-1}(x)|_0 = 2$. We are then equipped to give our next definition:

Definition 3. A Higher Order Efficient Bounded Tilting Estimator (EBTE) is a Bounded Tilting Estimator with bounded tilting divergence function q_{b^*} that also satisfies:

a) q_{b^*} is four times continuously differentiable with Lipschitz fourth derivative in a neighborhood of zero

b) $\frac{\partial^2}{\partial x^2} \dot{q}_{b^*}^{-1}(x)|_0 = 2$

By our choice of q_{b^*} we are then immediately granted with the main theorem of this section:

Theorem 6. *Under assumptions 1, 2, and 3, the EBTE are higher order efficient and when Y_θ is absolutely continuous with respect to a σ-finite measure for all θ, M_θ exists for all θ for all bounded tilting divergence functions q_{b^*}*

We have then arrived at a subclass of the MD estimators which deal with misspecification without any additional assumptions on the DGP and are higher order efficient.

2.4 Suggested EBTE and Simulations

2.4.1 Suggested EBTE

We now give an example of a tilting function $\dot{q}^{-1}$ which gives a member in this subclass of EBTE. The tilting function must be strictly increasing and bounded. To find such a candidate it makes sense to look at CDFs over the entire real line, most of which also satisfy the necessary Lipschitz differentiability criterion. However, the candidate must also satisfy the necessary order conditions $q^*(0) = 0$, $\frac{d}{dx}q^*(0) = \frac{d^2}{dx^2}q^*(0) = 1$, and $\frac{d^3}{dx^3}q^*(0) = 2$ which translate to $\dot{q}^{-1}(0) = \frac{d}{dx}\dot{q}^{-1}(0) = 1$ and $\frac{d^2}{dx^2}\dot{q}^{-1}(0) = 2$ by Lemma 1. $q^*(0) = 0$ is inconsequential because we simply normalize $\int \dot{q}^{-1}(x)dx|_{x=0} = 0$. This leaves us with 3 equations which must hold simultaneously. A simple family of CDFs which has a sufficient number of adjustable parameters is the combination of the Cauchy and the Logistic CDFs in the following manner:

$$\dot{q}^{-1}(x) = 2 * (\frac{1}{\pi}arctan(\frac{x-a}{b}) + \frac{1}{2} + \frac{1}{1+e^{-\frac{x-c}{d}}})$$

With $b = \frac{2\sqrt{6+5\sqrt{3}\pi}-3}{10\pi}$, $d = \frac{2(2\sqrt{3}\pi+\sqrt{6+5\sqrt{3}\pi})}{9(\sqrt{3}\pi-2)}$, $a = \sqrt{3}b$, and $c = ln(2)d$.

This yields the tilting function of an estimator in the EBTE class which we will call the Cauchy-Logistic EBTE.

2.4.2 Monte Carlo Simulations

We will now explore how our suggested example behaves in simulations. We will explore two setups in line with both Ragusa (2011) and Schennach (2007). The first setup explores the higher order efficiency of the proposed estimator, and is based on the design of Hall and Horowitz (1996). We use 13 sequences of independent random variables $\{\{x_{i,j}\}_{i=1}^n\}_{j=1}^{13}$ where:

$$(x_{i,1}, x_{i,2}) \sim_{i.i.d} \mathcal{N}([0,0], 0.16I), \quad x_{i,3} \sim_{i.i.d} t_5, \quad x_{i,j} \sim_{i.i.d} \chi^2_1 \;\; j = 4, ..., 13$$

We define $q(\theta, x, y) = exp(-0.72 - \theta(x + y) + 3y) - 1$ and implement four designs:

$$1) \quad g(X, \theta) = [q(\theta, X_1, X_2),\; X_2\, q(\theta, X_1, X_2)]'$$

$$2) \quad g(X, \theta) = [q(\theta, X_1, X_2),\; X_2\, q(\theta, X_1, X_2),\; X_3\, q(\theta, X_1, X_2),\; X_4\, q(\theta, X_1, X_2)]'$$

$$3) \quad g(X, \theta) = [q(\theta, X_1, X_2),\; X_2\, q(\theta, X_1, X_2),\; X_4\, q(\theta, X_1, X_2),\; ..., X_7\, q(\theta, X_1, X_2)]'$$

$$4) \quad g(X, \theta) = [q(\theta, X_1, X_2),\; X_2\, q(\theta, X_1, X_2),\; X_4\, q(\theta, X_1, X_2),\; ..., X_{13}\, q(\theta, X_1, X_2)]'$$

Design 2) is as suggested by Ragusa (2011) and Designs 3) and 4) are versions of what was done by Schennach (2007). Design 1) is symmetric whereas designs 2)-4) impose skewness and an increase in kurtosis. We implement 10000 replications for each design at $n = 100$ and $n = 400$. We discard the samples where the estimators failed to converge.[1] The results are shown in the tables below, together with the same results for EL and ET for comparison:

Table 2.1: Bias, Variance, Median, and Inter-Quartile Range (IQR) of EL, ET, and Cauchy-Logistic EBTE for $n = 100$ at 10000 replications (Design 1)

	Design 1): n=100		
Estimator	EL	ET	Cauchy-Logistic
Bias	0.056	0.072	0.063
Variance	0.083	0.088	0.086
Median	3.033	3.048	3.039
IQR	0.382	0.388	0.385

[1]This was detected by checking if providing three different starting points to the optimization routine gave different optima.

Table 2.2: Bias, Variance, Median, and Inter-Quartile Range (IQR) of EL, ET, and Cauchy-Logistic EBTE for $n = 100$ at 10000 replications (Design 2)

	Design 2): n=100		
Estimator	EL	ET	Cauchy-Logistic
Bias	0.113	0.181	0.149
Variance	0.093	0.118	0.107
Median	3.087	3.144	3.115
IQR	0.340	0.430	0.418

Table 2.3: Bias, Variance, Median, and Inter-Quartile Range (IQR) of EL, ET, and Cauchy-Logistic EBTE for $n = 100$ at 10000 replications (Design 3)

	Design 3): n=100		
Estimator	EL	ET	Cauchy-Logistic
Bias	0.162	0.267	0.226
Variance	0.103	0.147	0.131
Median	3.137	3.222	3.188
IQR	0.409	0.462	0.442

Table 2.4: Bias, Variance, Median, and Inter-Quartile Range (IQR) of EL, ET, and Cauchy-Logistic EBTE for $n = 100$ at 10000 replications (Design 4)

	Design 4): n=100		
Estimator	EL	ET	Cauchy-Logistic
Bias	0.284	0.450	0.399
Variance	0.127	0.213	0.182
Median	3.249	3.382	3.343
IQR	0.458	0.572	0.540

Table 2.5: Bias, Variance, Median, and Inter-Quartile Range (IQR) of EL, ET, and Cauchy-Logistic EBTE for $n = 400$ at 10000 replications (Design 1)

	Design 1): n=400		
Estimator	EL	ET	Cauchy-Logistic
Bias	0.019	0.022	0.019
Variance	0.020	0.020	0.020
Median	3.014	3.018	3.015
IQR	0.192	0.192	0.191

Table 2.6: Bias, Variance, Median, and Inter-Quartile Range (IQR) of EL, ET, and Cauchy-Logistic EBTE for $n = 400$ at 10000 replications (Design 2)

	Design 2): n=400		
Estimator	EL	ET	Cauchy-Logistic
Bias	0.031	0.053	0.039
Variance	0.020	0.021	0.021
Median	3.024	3.045	3.032
IQR	0.193	0.196	0.195

Table 2.7: Bias, Variance, Median, and Inter-Quartile Range (IQR) of EL, ET, and Cauchy-Logistic EBTE for $n = 400$ at 10000 replications (Design 3)

	Design 3): n=400		
Estimator	EL	ET	Cauchy-Logistic
Bias	0.046	0.085	0.064
Variance	0.021	0.023	0.022
Median	3.043	3.080	3.059
IQR	0.190	0.201	0.196

Table 2.8: Bias, Variance, Median, and Inter-Quartile Range (IQR) of EL, ET, and Cauchy-Logistic EBTE for $n = 400$ at 10000 replications (Design 4)

	Design 4): n=400		
Estimator	EL	ET	Cauchy-Logistic
Bias	0.084	0.156	0.123
Variance	0.021	0.026	0.024
Median	3.079	3.148	3.115
IQR	0.195	0.216	0.206

We can see that our Cauchy-Logistic EBTE performs more in line with the EL estimator than the ET with the difference between the estimators performance decreasing significantly with sample size as expected. As we move from Design 1) and 2) to designs 3) and 4) we see all three estimators' performing worse as a result of the increased skewness and kurtosis in the $n = 100$ designs. This effect is much less noticeable in the $n = 400$ designs but still there nevertheless. Designs 1) and 2) were based on Ragusa (2011) and the results are very similar to his for both the EL and the ET estimators while at the same time the Cauchy-Logistic EBTE performs approximately as well as Ragusa's Quartic Tilting estimator.

Our second simulation setup is intended to test how the estimators deal with misspecification. This setup was also used by both Ragusa and Schennach with moment function:

$$g(x, \theta) = [x - \theta, (x - \theta)^2 - 1]'$$

Our random variable x_i follows either a correctly specified model (we call this model C) or a misspecified model (we call this model M):

$$C) \quad \{x_i\}_{i=1}^n \sim_{i.i.d} \mathcal{N}(0,1)$$

$$M) \quad \{x_i\}_{i=1}^n \sim_{i.i.d} \mathcal{N}(0,0.64)$$

To see how the estimator using our Cauchy-Logistic tilting function perform we once again compare it to EL and ET. Our simulation entails first 10000 replications of $n = 1000$, followed by 2000 replications of $n = 5000$. The results are depicted in tables 2.9 and 2.10 respectively, where the standard deviations of the estimators are shown.

Table 2.9: Standard deviations of EL, ET, and Cauchy-Logistic EBTE estimators for Models C and M defined in the text with $n = 1000$

	n=1000		
Estimator	EL	ET	Cauchy-Logistic
Model C	0.0315	0.0315	0.0315
Model M	0.0547	0.0308	0.0280

Table 2.10: Standard deviations of EL, ET, and Cauchy-Logistic EBTE estimators for Models C and M defined in the text with $n = 5000$

	n=5000		
Estimator	EL	ET	Cauchy-Logistic
Model C	0.0144	0.0144	0.0144
Model M	0.0520	0.0140	0.0123

Our Cauchy-Logistic EBTE performs as well as the other estimators in the well specified model, as is also highlighted in the previous simulation setup. As predicted, in the misspecified model our EBTE performs even better than ET which is well equipped to handle misspecification. This suggests that, even when the DGP is such that ET admits a pseudo-true value, there may still be further benefits to using a bounded tilting func-

tion. Note finally that just like Schennach (2007) we find that the EL variance in the misspecified model is significantly larger than the others and also does not decrease with sample size by the expected $\sqrt{5}$ factor like the others do.

2.5 Conclusion

Since the introduction of Hansen's GMM there have been ample resources devoted to finding ways to improve upon the techniques and the ECR estimators did so in many ways. Unfortunately the search for the "best member" of this subclass of MD estimators has been misguided in that there are no members which can satisfy the two requirements of robustness to misspecification and higher order efficiency simultaneously. Even then, previous suggestions of ways to get around this issue have come with the requirement of heavy assumptions on the DGP and moment condition function g. Our approach of looking at a different subclass of the MD estimators which have a bounded tilting function eliminates the need for any assumptions to guarantee a solution to the asymptotic tilting problem and, by narrowing our scope to members which satisfy a derivative condition, we produce the class of EBTE which is also higher order efficient. This effectively renders the class of ECR estimators obsolete and finishes the search for GMM's replacement according to the two criteria we have followed in this chapter.

A natural way to produce bounded tilting functions is to use combinations of scaled CDFs and we provide an example which we dub the Cauchy-Logistic EBTE. We simulate the Cauchy-Logistic EBTE together with EL and ET in first a well specified setting with varying degrees of skewness and kurtosis. The Cauchy-Logistic EBTE performs more in line with EL than ET as suggested by its higher order efficiency. Secondly we impose a misspecified model, where our simulations show that the Cauchy-Logistic EBTE outperforms both EL and ET.

While the EBTE class completes the search for the "best estimator" according to the

two strands of research we have mentioned, we hope that this may pave the way for future research and hence we leave the reader with an open question. Since these estimators are both first order and higher order efficient while at the same time ensuring robustness to misspecification with minimal assumptions, can we further narrow down the class to yield even more optimality or implementability properties?

CHAPTER 3

RESPONSES TO INFORMATION OBFUSCATION IN THE LABORATORY

3.1 Introduction

There are many real-life settings in which a principal is concerned with the amount of effort an agent exerts on a task of difficulty ex-ante unknown to both the principal and the agent, and where the principal can only motivate the agent with information about the difficulty rather than monetary rewards. Consider as an example a college professor and an undergraduate student. The professor wants the student to work hard to reach the objectives of her course, while the student only cares about passing the course while exerting as little effort as possible. There might be long term benefits that the student receives from learning as much as possible. However, we assume that these are not realized until a later point in time and, hence, the incentives of the professor and student are not aligned in this particular situation. When the course starts it is not clear to the professor what constitutes a passing grade since the quality distribution of her students has not yet been revealed. However, after the first midterm or homework, the professor will have more information and can choose how much of this to communicate to the students. Since the effort provision is continuous throughout the course, revealing perfectly the threshold necessary to pass enables the student to reduce the amount of effort she provides to just pass.

Ely and Szydlowski (2020) study this problem and derive a theoretically optimal information revelation strategy for the principal. The key to this strategy is not only *what* signal to send but also *when* to send it - precisely because time passes as effort is provided. The prescribed strategy for when an agent is sufficiently optimistic about task difficulty to start exerting effort - or in our example, to enroll in the course - is dubbed by Ely and

Szydlowski (2020) as "leading the agent on" (LEAD)[1]. We restrict ourselves to the binary case where the difficulty level can only be high or low, leading to a high or low amount of effort necessary for the agent to complete the task respectively. In this case, the LEAD strategy has the principal committing to truthfully send a signal according to an ex-ante specified probability distribution about the true difficulty level *after* the agent has exerted enough effort to have finished the low difficulty task. This signal truthfully reports high difficulty but misreports the low difficulty as high difficulty just often enough to make the agent indifferent between stopping and continuing with the effort provision until the difficult task is complete. The aspects of committing ex-ante to a signal structure and revealing truthfully the realization to the agent link this work to the growing literature of Bayesian Persuasion. The term was coined by Kamenica and Gentzkow (2011) but the main ideas were first introduced by Aumann et al. (1995) in their work on repeated games with incomplete information.

The optimality of the LEAD strategy relies on the receiving agent maximizing her own financial interest in a rational manner. However, it is a well-established finding across all strands of the experimental economics literature that observed behavior deviates from the perfectly rational homo economicus. Consequently, in this project, we explore the setting laid out above through the use of a discretized version of the model from Ely and Szydlowski (2020) in the laboratory. The principal, which we refer to as the Sender in our experiment, chooses an information structure which is implemented by a computer and the agent, which we refer to as Receiver, needs to respond to this information[2]. We abstract away from real effort in the experiment in that responding to the information is done by choosing how many consecutive periods to participate in. Choosing not to participate in a given period ends the interaction between the Sender and Receiver. The

[1]When the agent is not sufficiently optimistic about the task difficulty, the optimal strategy derived by Ely and Szydlowski (2020) is dubbed "Moving the Goalposts" and requires an additional signal to be sent before the agent starts providing effort.

[2]We use these expressions to conform with the existing experimental literature on Bayesian Persuasion. See, for example, Fréchette et al. (2019) and Nguyen (2017).

Receiver receives a reward if she participates up to and including a threshold period. Theoretically this problem is equivalent to the effort provision setup but the framing is more general.

As mentioned above, our interest is in whether the agents are rational and profit maximizing which we explore in our experimental setting through the responses of the Receivers. For this purpose we do not allow the Senders to freely choose how and when to send information. Since we are also interested in the optimality of the LEAD strategy we do not restrict the Senders to only choose the LEAD strategy but rather give them a discrete list of information structures to choose from. This list consists of the LEAD strategy along with three other focal policies: full information about the threshold period before the agent has to make any decisions, no information about the threshold period at any point, and a strategy that keeps the same timing of the LEAD strategy but reveals the true threshold period fully rather than misreporting sometimes[3].

Our experimental results show that compliance, i.e. optimal theoretical response to a given information structure, varies with the informational content provided in the information structures[4]. The compliance rate is increasing with information. This behavior actually leads the LEAD strategy to be outperformed by providing the Receiver with full information about the threshold period after the Receiver has participated in the number of periods that corresponds to the low threshold period. In our model section and in calculating the optimal responses, we are assuming that the subjects are risk-neutral, just as Ely and Szydlowski (2020). However, allowing for some degree of risk-aversion yields the theoretical prediction that the delayed full information strategy outperforms the LEAD strategy. Hence, subjects exhibiting a range of different risk-attitudes is a feasible explanation for this observation.

[3]To be precise, we are only using approximately the optimal LEAD strategy. This is because the optimal policy makes the agent indifferent between continuing the effort provision and stopping while our version makes continuing strictly preferable.

[4]Throughout the chapter, we will use the expression "informational content" to not only express the amount but also the timing of the information. Hence, an information structure that provides the same information amount as another but at a later point is considered to have lower information content.

Moreover, we find that when Receivers receive full information about the threshold period before having to make participation choices, they participate on average in more periods than theoretically predicted. Receiving information through any of the other three information structures leads to a lower than optimal average participation, where the gap between actual and optimal average participation is increasing in the amount of uncertainty contained in the information structure. Reciprocity would perfectly explain this finding. The LEAD strategy, as the name suggests, involves the agent being led on. That means that the principal withholds at first and then obfuscates information to prompt the agent to exert as much effort as possible for the principal's personal gain. Stringing an agent along for the principal's own benefit might lead the agent to forgo some own benefit to hurt the principal. In experimental labor markets it has been shown that the agent punishes the principal if a suggested wage rate is perceived as unfair. Examples include Fehr et al. (1998), Charness and Rabin (2002), and Charness (2004). Charness (2004) establishes a positive relationship between effort and wage. Additionally, he shows that for low wage rates the average effort is lower when the wage rate was chosen by a subject employer rather than a random process. This suggests that in our setting, leading the agent on might provoke a similar response of negative reciprocity from the agent. At the same time, presenting the agent with as much information as possible might make the agent willing to reciprocate and exert some additional, unprofitable effort. This is supported by findings in the same paper, Charness (2004): For high wage rates, the average effort is higher when the wage rate was chosen by a subject employer rather than a random process.

Hence, our second set of findings is in regards to whether the non-compliance with the theoretical optimal response to the information structures can be explained by the responses of the Receivers exhibiting reciprocity. To our knowledge we are the first to analyze reciprocity of an agent towards a principal when the principal reveals information

rather than sets a wage in an effort-provision setting[5]. The latter has been extensively studied in the experimental literature in the labor market context. For example, Fehr et al. (1998) simulate labor markets in the lab with some subjects taking on the role of the employer and others the role of the employee. The authors compare effort provision when wages are determined by employers or by an external process. In line with this research, in our experiment the information structure chosen by the Sender is only implemented half of the time, while the other half of the time the computer randomizes which information structure is implemented. The Receiver is informed about which information structure was chosen and whether it was chosen by the Sender or the computer. Our experimental results show that when the Sender is generous with the information and chooses to reveal the true threshold period before the Receiver has to make any choices, then the Receiver participates in more periods on average than when it was chosen randomly by the computer. While this can be interpreted as positive reciprocity towards the Sender, this result is not robust to outliers and can be explained by two subjects making mistakes. For the remaining information structures, the findings are not in line with the experimental labor market with the Receivers participating in weakly more periods on average for every structure and every message when it was chosen by the Sender compared to when it was chosen by the computer.

At its core, our experiment is concerned with how to motivate effort provision in the lab. We have found that delayed information revelation leads to higher effort provision than revealing all the information before an agent starts exerting effort or not revealing any information at all. Whether full or obfuscated information is provided after a delay does not have an impact on the effort exerted in our experiment. As such our project contributes and is related to the literature on this topic. DellaVigna and Pope (2018) study factors that motivate effort provision when effort is rewarded with a flat payment. Subjects receive a flat payment no matter how much effort they exert and are incentivized

[5]Reciprocity in relationship to information has been studied before but not in the effort-provision setting. Au and Li (2018) study Bayesian Persuasion in the presence of reciprocity.

by additional monetary payoffs. The treatment closest to our experiment is one where subjects are paid a bonus for completing a certain number of tasks. However, the task is such that subjects can easily infer how many tasks they have completed correctly. Eriksson et al. (2009) compare effort provision under different feedback rules. However, all information provision is in terms of relative performance compared to an opponent and only the best performing agent receives a prize. Chen and Schildberg-Hörisch (2018) also vary amount of feedback provided. They consider only individual performance but test how the information impacts effort provision in a post-revelation effort task compensated in a piece-rate fashion.

The remainder of this chapter is organized as follows: Section 1.2 revisits Ely and Szydlowski (2020) and adapts their binary framework for experimental analysis by discretizing effort provision. Section 1.3 outlines the experimental design and the implementation in the laboratory. It also includes the hypotheses we are looking to test in our experiment. Section 1.4 presents and discusses the results of the experiment. Section 1.5 concludes. Screenshots and complete experimental instructions are included in the appendix to this chapter, section 1.6.

3.2 Model

3.2.1 Setup

Just as in Ely and Szydlowski (2020), we consider an agent who works on behalf of a principal. The agent spends effort e on her work in increments until the agent chooses to quit. In Ely and Szydlowski (2020) the increments are continuous making the variable e a continuous variable whereas we will discretize and normalize so that the incremental effort provision is 1 unit[6]. The effort provided by the agent produces an output $y(p, e)$ which

[6]You can think of this as there being effort choices e_t for each discrete period t. The agent then in each period decides whether $e_t = 1$ or $e_t = 0$. If $e_t = 0$ is ever chosen the game ends. The total effort e is then $e = \sum_{t=0}^{s} e_t$ with s being the period in which the agent stops

is determined by her productivity level p and her effort level. While Ely and Szydlowski (2020) normalize the productivity parameter to 1 - something which we will do later - it is useful to see the effect it has on the derivations. For simplicity we assume a production function of the form $y(p, e) = pe$. The effort is also costly to the agent in that each unit of effort provided costs c. When the agent chooses to stop providing effort the agent is rewarded with a payoff $R > 0$ if and only she has produced more than a threshold $x > 0$ (i.e. if $y(p, e) \geq x$).

As in Ely and Szydlowski (2020) the threshold x is unknown to the agent which induces uncertainty about whether the task has been successfully completed or not at a given effort level. We formalize this uncertainty by giving the agent a prior with CDF F over the unknown $x \in X \subset \mathbb{Z}_+$, where X is the space of possible thresholds[7]. To make the setup easier to understand for our subjects, just like Ely and Szydlowski (2020) do in the introductory part of their paper, we limit ourselves to a binary set-up where $X = \{x_l, x_h\}$, where x_l indicates a low and x_h a high threshold. We can summarize the beliefs about the true success threshold using a parameter μ instead of the CDF F:

$$x = \begin{cases} x_l & \text{with probability} \quad \mu \\ x_h & \text{with probability} \quad 1 - \mu \end{cases} \tag{3.1}$$

For simplicity, and for applicability in our experimental design, we also do not include discounting in this model. We also make the innocuous addition that we endow the agent with a budget that she can use to pay for effort with and hence we write the agents expected payoff from providing an effort level e as:

$$\pi_a(e) = \mathbb{E}[\mathbf{1}[x \leq pe]R - ec] = B + F(pe)R - ec \tag{3.2}$$

The principal does not pay the cost of the agents reward but still receives a benefit

[7]Since we have discretized the effort levels e we may also discretize the space X and let it take only whole positive numbers.

$r > 0$ from each unit produced by the agent. We can hence write the principal's expected payoff as:

$$\pi_p(e) = rpe \tag{3.3}$$

While the agent does not know the threshold x, the principal does and can hence choose what information to disclose and when to disclose it. However, the principal must commit to a revelation policy before learning the threshold and we assume, just like in Ely and Szydlowski (2020), that the agent knows the policy and understands the principal's commitment. The theoretical work in Ely and Szydlowski (2020) is then done to produce an "optimal" revelation policy for the principal - the policy which maximizes the payoff of the principal. As described in the introduction one goal of this chapter is to see whether the optimal policy derived in Ely and Szydlowski (2020) works as predicted in a lab setting. To derive the optimal information policy we must first understand the incentives of the agent. We can then write the agent's expected payoff function as:

$$\pi_a(e) = \begin{cases} B - ec & if\, e < \frac{x_l}{p} \\ B + \mu R - ec & if\, \frac{x_l}{p} \leq e < \frac{x_h}{p} \\ B + R - ec & if\, \frac{x_h}{p} \leq e \end{cases} \tag{3.4}$$

Since in all three cases incremental units of effort only decreases the payoff unless it moves you into a higher case, there are only three possible rational options for the agent to make in her effort choice. These are $e = \{0, \frac{x_l}{p}, \frac{x_h}{p}\}$ which lead to expected payoffs $\pi_a^\mu = \{B, B + \mu R - \frac{x_l c}{p}, B + R - \frac{x_h c}{p}\}$ respectively.

We now make the same assumptions as those in Ely and Szydlowski (2020), the first of which will ensure that working until x_h units are produced without any additional information is never rational but working until x_l units can be rational.

Assumption 4. $\frac{x_l}{p} < \frac{R}{c} < \frac{x_h}{p}$

We also maintain the assumption that the task x_h is not too difficult by assuming that:

Assumption 5. $\frac{x_h - x_l}{p} < \frac{R}{c}$

We summarize what these two assumptions give us in the following lemma:

Lemma 3. *Assumption 1 implies that there is no μ for which it is rational to work until x_h units are produced but also that $\exists\ \bar{\mu} \in (0,1)$ for which the agent will work until x_l units are produced if $\mu \geq \bar{\mu}$.*

Assumption 2 implies that an agent who has worked until x_l units are produced will find it profitable to work until x_h units are produced if she is told with certainty that $X = x_h$.

Proof. For the Assumption 1 implication, consider the "best case" for x_h where the agent knows with certainty that it is the true threshold ($\mu = 0$). This leads to a choice between $\pi_a^0 = \{B, B - \frac{x_l c}{p}, B + R - \frac{x_h c}{p}\}$. Clearly choosing $\frac{x_h}{p}$ strictly dominates $\frac{x_l}{p}$ but under assumption 1, 0 also strictly dominates $\frac{x_h}{p}$ since:

$$\frac{R}{c} < \frac{x_h}{p} \iff R - \frac{x_h c}{p} < 0 \iff B + R - \frac{x_h c}{p} < B \iff \pi_a^0(\frac{x_h}{p}) < \pi_a^0(0)$$

Consider then the tradeoff between choices 0 and $\frac{x_l}{p}$. Because of the slack in the assumption there must exist a $\bar{\mu}$ such that for all $\mu > \bar{\mu}$ it is rational to start working and work until $\frac{x_l}{p}$ units are provided. We find this $\bar{\mu}$ by solving:

$$\bar{\mu} R - c\frac{x_l}{p} = 0,$$

which gives us

$$\bar{\mu} = c\frac{x_l}{pR}$$

For the Assumption 2 implication, consider the tradeoff between $\frac{x_h}{p}$ and $\frac{x_l}{p}$ when $\mu = 0$:

$$\frac{x_h - x_l}{p} < \frac{R}{c} \iff B - \frac{x_l c}{p} < B + R - \frac{x_h c}{p} \iff \pi_a^0(\frac{x_l}{p}) < \pi_a^0(\frac{x_h}{p})$$

□

3.2.2 Optimal revelation policy

In deriving the optimal revelation policy, Ely and Szydlowski (2020) consider the effect of using information as a carrot to get the agent to produce effort. Telling the agent that she will receive information about the true state of the world after producing x_l units will be an incentive in itself. To investigate this, we first define the benefit to the agent of full information disclosure after $\frac{x_l}{p}$ units of effort:

$$V(\mu) = \mu R + (1 - \mu)[R - c\frac{x_h - x_l}{p}]$$

With probability μ the agent is told that the true state is x_l, stops working, and receives benefit R and with probability $1 - \mu$ the agent is told that the true state is x_h and keeps working until then because of assumption 2.

The cost to pay for this information is that the agent must work until $\frac{x_l}{p}$ effort has been provided and she must hence give up $c\frac{x_l}{p}$. The range of priors for which the agent is willing to provide $\frac{x_l}{p}$ units of effort and get the full information revelation is hence:

$$V(\mu) - c\frac{x_l}{p} \geq 0$$

We may hence define $\tilde{\mu}$, the smallest value of μ for which the agents is willing provide $\frac{x_l}{p}$ units of effort to receive the information as:

$$R - (1 - \tilde{\mu})(c\frac{x_h - x_l}{p}) - c\frac{x_l}{p} = 0$$

Or after simplifying:

$$\tilde{\mu} = \frac{x_h - \frac{pR}{c}}{x_h - x_l}$$

Note that given assumption 2, we have that $\bar{\mu} > \tilde{\mu}$.

Ely and Szydlowski (2020) prove that if $\mu > \tilde{\mu}$ the the principal can do better than fully revealing the state after an effort of $\frac{x_l}{p}$ has been provided by sending a signal instead. They call the information structure which the principal optimally employs the "leading the agent on" (LEAD) strategy. For the remainder of this chapter we will focus exclusively on this case by assuming that:

Assumption 6. $\mu > \tilde{\mu}$

Since there is slack in assumption 6, the optimal strategy involves telling the agent that she will receive a signal s after x_l units have been produced. This signal will extract as much of the agent's surplus as is possible. The probability structure of this signal is:

Table 3.1: Signal structure

	$x = x_l$	$x = x_h$
$s = x_l$	$1 - q$	0
$s = x_h$	q	1

The signal s discloses the true state x_h if x_h is the true state, and says that the state is x_h with probability q and x_l with probability $1 - q$ if the true state is x_l. After receiving a signal of x_l, an agent knows with certainty that the true success threshold is x_l and will stop working. After receiving a signal of x_h, the agent does not with certainty know what the state is. If we can ensure that the agent complies with the signal's outcome (provides $\frac{x_h}{p}$ if the signal outcome is x_h) then the principal can extract more effort from the agent than under full information by making q as large as possible.

We will hence need two familiar conditions to hold, an Individual Rationality constraint (IR) and an Incentive Compatibility constraint (IC). The IC constraint will ensure that

the agent complies with the signal outcome. The IR constraint will ensure that it is (weakly) profitable for the agent to start providing effort and go until $\frac{x_l}{p}$ to receive the information given that the IC constraint holds.

Starting with the IC constraint, the agent will always comply with the signal if the signal realization is x_l since this implies that the state is x_l with probability 1. However the reason for why an IC constraint is necessary is that the agent will not always comply with the signal if the signal realization is x_h. The trade off for the agent is between receiving $B + R - \frac{x_h c}{p}$ from working until $\frac{x_h}{p}$ and receiving $B + R(1 - P(x_h|s = x_h)) - \frac{x_l c}{p}$ from staying at $\frac{x_l}{p}$. This means that in order for the agent to comply we need:

$$B + R - \frac{x_h c}{p} > B + R(1 - P(x_h|s = x_h)) - \frac{x_l c}{p}$$

Which we may rewrite to:

$$P(x_h|s = x_h) > c\frac{x_h - x_l}{pR}$$

Writing this in terms of q and μ we get since $P(x_h|s = x_h) = \frac{1-\mu}{1-\mu+q\mu}$:

$$\frac{1 - \mu}{1 - \mu + q\mu} > c\frac{x_h - x_l}{pR}$$

Simplifying this we get a constraint on q:

$$q < [\frac{1 - \mu}{\mu}][\frac{1 - c\frac{x_h - x_l}{pR}}{c\frac{x_h - x_l}{pR}}] = q^*_{IC} \tag{3.5}$$

Given the above IC constraint we can now evaluate the IR constraint. Notice that should the IC constraint not hold, then the agent may choose not to work until $\frac{x_h}{p}$ upon a signal realization of x_h which would change the trade off. Now, with this signal structure the agent knows the probabilities of the signal realizations and hence the willingness to start working at 0 effort provided is determined by a different condition on μ than when

full information was provided. The agent knows that with probability $\mu_p = (1-q)\mu$, the probability of a signal realization of x_l, she will stop at $\frac{x_l}{p}$ and the condition is hence:

$$V(\mu_p) - c\frac{x_l}{p} \geq 0$$

Which becomes the following condition on q:

$$q < \frac{R - c\frac{x_h}{p} + \mu c\frac{x_h - x_l}{p}}{\mu c\frac{x_h - x_l}{p}} = q^*_{IR} \tag{3.6}$$

Combining the IC and IR conditions we can show the following:

Proposition 7. Under assumptions 4, 5, and 6 and if:

i) $\mu < \bar{\mu}$

Then the IR constraint (3.6) is the binding inequality and hence the optimal distortion level q^* is q^*_{IR}. This leads to the agent's ex-ante utility being set to zero. The agent completes the task with probability one.

ii) $\mu > \bar{\mu}$

Then the IC constraint (3.5) is the binding inequality and hence the optimal distortion level q^* is q^*_{IC}. This does not lead to the agent's ex-ante utility being set to zero but the agent still completes the task with probability one.

Proof. We start by comparing q^*_{IC} and q^*_{IR}:

$$\frac{q^*_{IR}}{q^*_{IC}} = \frac{\frac{R - c\frac{x_h}{p} + \mu c\frac{x_h - x_l}{p}}{\mu c\frac{x_h - x_l}{p}}}{[\frac{1-\mu}{\mu}][\frac{1 - c\frac{x_h - x_l}{pR}}{c\frac{x_h - x_l}{pR}}]} = \frac{1 - c\frac{x_h}{pR} + \mu c\frac{x_h - x_l}{pR}}{(1-\mu)(1 - c\frac{x_h - x_l}{pR})} \tag{3.7}$$

Working with the denominator we can see that:

$$1 - c\frac{x_h}{pR} + \mu c\frac{x_h - x_l}{pR} = (1-\mu)(1 - c\frac{x_h - x_l}{pR}) - c\frac{x_l}{pR} + \mu$$

Plugging this into (3.7) we get:

$$\frac{q^*_{IR}}{q^*_{IC}} = \frac{(1-\mu)(1-c\frac{x_h-x_l}{pR}) - c\frac{x_l}{pR} + \mu}{(1-\mu)(1-c\frac{x_h-x_l}{pR})}$$

Notice that $1 - c\frac{x_h-x_l}{pR} > 0$ by our assumption 5, which means that the fraction is clearly bigger than 1 if $-c\frac{x_l}{pR} + \mu > 0$. In other words:

$$q^*_{IR} > q^*_{IC} \iff -c\frac{x_l}{pR} + \mu > 0 \iff \mu > \bar{\mu} \tag{3.8}$$

Since both the IR and IC constraint must hold it must be that $q^* = q^*_{IC}$ iff $\mu > \bar{\mu}$.

When we have $\mu < \bar{\mu}$ then q^* is optimal in that it extracts all the agent's surplus so there cannot exist a better strategy for the principal. This is because the IR constraint binding equates to the agent being indifferent between getting 0 and taking part of the mechanism.

When we have $\mu > \bar{\mu}$ the agent will still comply with the signal outcome and hence complete the task with probability one. There cannot exist a better information revelation strategy for the principal because once x_l has been reached we have maximized the probability that the agent continues to $\frac{x_h}{p}$. Sending a signal at time 0 which tells the truth about x_h would make it so that the agent will sometimes not start and hence not complete the task with probability 1 which means that we cannot do better than using q^*. Telling the truth about x_l means that we cannot make the agent more pessimistic about x_l and hence we can not improve our q^*. □

This proposition agrees with Ely and Szydlowski (2020) for $\mu < \bar{\mu}$ but when $\mu > \bar{\mu}$ the incentive constraint binds before the IR constraint and hence the proposition disagrees with it in that you cannot extract all the surplus from the agent. A comment must be made here in regards to the Ely and Szydlowski (2020) model in which the authors do not consider an IC constraint. It is not clear whether such a constraint is necessary in their setup but the logic above makes it clear why we need it here. Perhaps further inspection

of the model of Ely and Szydlowski (2020) with this IC constraint in mind is necessary.

3.3 Experimental design and implementation

Section 3.2 presented the optimal information structure, denoted by LEAD, in a discretized version of the binary setting of Ely and Szydlowski (2020). The optimality of the information structure depends on the agents responding to it in the theoretically optimal way. However, as mentioned in the introduction, it is not clear how well this strategy will perform in the laboratory with subjects responding to it. Hence, we design a laboratory experiment to test how agents respond to being lead on and to other focal information structures about the threshold x. We further design it to test whether attribution influences the amount of effort invested into a task. We will now formalize these goals into testable hypotheses and discuss our design choices.

In the implementation, we use a version of stated effort. This is because in order to define the optimal information revelation as elaborated on in section 3.2, we require a constant and known cost of effort and in the "stated-effort approach [...] there is no uncertainty regarding an individual's cost of effort" (Charness et al., 2018). Subjects have to decide how many consecutive periods to participate in by, starting in the first period, choosing to continue or stop participating. Choosing to participate incurs a cost of $c = \$1$ that is deducted from a budget of $B = \$10$ that each subject is endowed with. Subjects will win a reward $R = \$8$ if they participate for sufficiently many periods. How many periods are sufficient is randomly determined before subjects begin making participation choices. It is either $x_l = 5$ periods with probability $\mu = \frac{2}{3}$ or $x_h = 9$ periods with probability $1 - \mu = \frac{1}{3}$. In this way we create a common prior for all the subjects. We call the last period in which participation is necessary to win the reward the threshold period. In choosing this set-up, we are implicitly setting $p = 1$[8]. The Sender earns $r = \$1.50$

[8]Note that this is another reason for not opting for subjects to complete real-effort tasks. The productivity of subjects would have varied, once again making it impossible to calculate the optimal

for each period the Receiver participates in. The numbers have been chosen to conform with all the assumptions made in section 3.2. Theoretically this problem is equivalent to the effort provision setup but is general enough so as to not influence subjects' decisions by creating associations. Also, we will rename the principal to "Sender" and the agent to "Receiver" in our implementation to conform with existing experimental literature on information design.

As outlined in the introduction, our interest lies in the Receivers' behavior so we do not care how Senders choose information structures and, hence, do not allow them to freely choose how and when to send information. Instead, we provide the Senders with the following discrete list of strategies. Limiting the choice of the Senders to a list is attractive from an implementation standpoint due to findings from Fréchette et al. (2019). In a laboratory setting they test, amongst other things, Bayesian Persuasion and show that the Sender behavior is very heterogeneous, with a large group of subjects over-communicating and another large group of subjects under-communicating. Since even in their one-dimensional setting there is a wide range of chosen signal structures, adding the time dimension would likely result in too much heterogeneity to statistically compare the performances of the different structures. The following is the list chosen for our implementation:

Information Structures:

A - Full information before period 1

B - No information

C - Partial information after 5 periods

D - Full information after 5 periods

Information structures A and D can be seen as benchmarks in that they illustrate the benefits of just delaying information as opposed to both delaying and obfuscating as in C.

information revelation.

The Sender obfuscates information in C by reporting a signal realization. The underlying signal structure is depicted in table 3.1 in section 3.2, and setting $q = q^*$ is the LEAD strategy. For our implementation we choose $q = \frac{2}{3}$, which is only approximately equal to LEAD. Setting $q = q^*$ would make the subject exactly indifferent between continuing to 9 and stopping participation. We allow for some slack so that continuing is strictly better than stopping after 5 periods. This choice of q satisfies both our IC and IR constraint as defined in equations (3.5) and (3.6). Note that when we explain information structure C to the subjects in the instructions (see section 1.6.2), we give the updated probabilities of each state after receiving the message "The threshold period is 9." The reason we do this is that information structure C is difficult as it is. Receivers are informed about the timing and the signal structure. Then, based on their prior they have to make a choice about whether to start participating or not. Finally, they have to use Bayesian updating after receiving the signal, if they have chosen to participate long enough to receive the message, and use this updated information to make a choice about whether to continue participating or stop. Experimental papers including Tversky and Kahneman (1973) and Charness and Levin (2005) provide convincing evidence that subjects in the laboratory have difficulty using Bayesian updating correctly[9]. We are not concerned with this and, hence, reduce the complexity of the structure by an aspect that we already know subjects are not good at. As a final remark, B is included to not make D seem like an acceptable compromise.

The risk neutral optimal response by the agents to these informations structures is as follows:

Optimal Response:

A - Participate for 5 periods if the threshold period is revealed to be 5 and participate for 0 periods if it is revealed to be 9

B - Always participate in 5 periods

[9] A more extensive list can be found in the introduction of the latter paper.

C - Participate for 5 periods if the signal realization is 5 and participate for 9 periods if it is 9

D - Participate for 5 periods if the threshold period is revealed to be 5 and participate for 9 periods if it is revealed to be 9

We are now equipped to formulate our first goal of the experiment into a testable hypothesis:

Hypothesis 1. *(Compliance) Receivers respond optimally to the choice of information structure as defined above.*

A plausible reason for hypothesis 1 not to hold is positive and negative reciprocity[10]. As mentioned in the introduction, in experimental labor markets it has been shown, for example by Charness (2004), that agents punish their principals if they suggest a wage rate that is perceived as unfair and reward their principals by additional effort provision if they suggest a wage rate that is perceived as generous. This behavior was in comparison to the benchmark where the wage rate was randomly drawn or set by the experimenter. While these effects of positive and negative reciprocity caused by attribution have so far only been shown when the principal could directly determine a payoff-relevant variable, we are interested to test whether this plays a role in participation choices of the Receiver in our setting when the Sender chooses when and how much information to send. To test this we need a comparison of Receiver behavior when the information structure is chosen either by a computer or by a participant Sender. In the implementation of our experiment we are limiting ourselves to one treatment where the Sender's choice is implemented with 50% probability and a random computer choice is implemented with 50% probability. This is different from Charness (2004) where in one treatment, the employers' choices are implemented and in the other treatment, a random process (or the experimenter)

[10]We shall adhere to Charness's definition of reciprocity in that it is "the degree to which an intentional choice by a self-interested party induces a change, relative to the same choice being made without that party's volition, in a responding party's willingness to sacrifice money to help her" Charness (2004).

determines the wage rate. When the wage is not determined by the employers, they do not have any impact on the experiment and exist for the sole purpose of receiving a payoff based on the effort choices of the employees, a problem which we eliminate with our setup as every Sender always has to make a choice that could be implemented. Of course, the Receiver is informed about whether the amount and timing of the information was selected by another subject in the role of the Sender or by the computer (see Figure 1.9 in section 1.6.1) so that we can see whether it has an impact. This allows us to test the following hypothesis:

Hypothesis 2. *(Reciprocity) Receivers respond with positive reciprocity when the Sender truthfully reveals the threshold period either before the Receiver needs to make choices or after 5 periods and Receivers respond with negative reciprocity when the Sender additionally obfuscates or never sends any information.*

Let us now provide some more detail about the actual implementation of our experiment in the laboratory. Each subject takes part in two Sender-Receiver interactions. In one interaction, the subject takes on the role of a Sender and is paired with another subject taking on the role of a Receiver and vice versa. A subject does not learn anything about the outcome from the interaction in which she was the Sender before having to make her choice as a Receiver. No participant faces the same other participant in both these interaction. While this setup allows us to only obtain one observation from each subject, we are guaranteed that learning or reacting to past experiences does not impact our data.

Each interaction consists first of an information stage, and then of a participation stage made up of ten periods. In the information stage (see figure 1.8 in section 1.6.1), the Sender has to choose one of the four information structures introduced above. After choosing, different than in the model introduced in section 3.2, the Sender does not learn the true threshold. This is because if the Sender's choice is selected to be the structure according to which the Receiver receives information, it is implemented by the computer

so that the Sender cannot make any mistakes in implementing it[11]. In the participation stage, the Receiver has to choose after how many periods to end participation and with it the interaction. The interaction ends either if the Receiver has spent her entire budget, i.e. participated in 10 periods, or if the she chooses not to participate in one of the ten periods. The Receiver is provided with all the necessary information to make her choice (see figures 1.10, 1.11, and 1.12 in section 1.6.1).

The true threshold period along with the number of periods that the Receiver has chosen to participate in determines the earnings for the Sender and Receiver for an interaction. The earnings functions are given by:

$$\pi_R = B + \mathbf{1}[x \leq e]R - ce$$

$$\pi_S = re$$

Subjects are paid the earnings of one of the two interactions, randomly determined, along with a $5 show-up fee. Detailed paper instructions (included in section 1.6.2) were handed out before the beginning of the experiment. Subjects kept and could refer to the instructions until the end of the experiment. After the instructions were read out, the subjects had to answer a series of understanding questions. Subjects could not move on to the next part of the experiment without answering all of these correctly. Subjects played 5 practice rounds before the start of the two payoff-relevant interactions. Each of these practice rounds also consisted of two interactions, but subjects played against themselves.

We ran 7 sessions in the Brown University Social Science Experimental Laboratory (BUSSEL) in March 2020[12]. Participants were students from Brown University and the Rhode Island School of Design (RISD), recruited through the BUSSEL website. A total of 126 subjects participated in the experiment. Each experimental session lasted between

[11]This is in accordance with Fréchette et al. (2019).

[12]We ran an 8th session but one of the subjects chose to leave the session while it was running. Consequently, we were not able to collect data from the 19 participants in that session.

45 minutes and one hour. Average earnings, including a $5 show-up fee, were $14.13. The experiment was programmed and conducted with the experiment software z-Tree (Fischbacher, 2007). Table 3.2 provides an overview of the data we have collected. The "Total" column provides the number of observations of each information structure/threshold period/message combination. In the "Sender" and "Computer" columns, these numbers are split up based on whether the info structure was chosen by the Sender or the computer. The rightmost column as well as "Sender total" and "Computer total" provide aggregate observation numbers for each information structure.

Table 3.2: Data overview

Info structure	Threshold period	Message	Sender	Sender total	Computer	Computer total	Total	
A	5	5	9	14	10	11	19	25
	9	9	5		1		6	
B	5	-	2	4	13	19	15	23
	9	-	2		6		8	
C	5	5	13	25	13	20	31	45
		9	3		2			
	9	9	9		5		14	
D	5	5	10	17	11	16	21	33
	9	9	7		5		12	
				60		66		126

3.4 Results and discussion

In this section we present the two main sets of results of our experiment in response to the two hypotheses laid out in section 1.3. The first set of results deals with how well subjects comply with the theoretically predicted behavior, as outlined in hypothesis 1. To answer this, we consider the data as a whole and pool the data from when the information structure is chosen by a Sender and by the computer. We also provide some insight on our findings through the lens of risk aversion. Our second set of results addresses the existence of reciprocity in the choices of the Receivers in response to hypothesis 2. To answer this

we split the data set into observations where the information structure was chosen by the Sender and where the information structure was chosen by the computer. Comparing these two subsets of our data allows us to see whether there is a causal attribution effect, and more specifically whether the choices exhibit reciprocity.

3.4.1 Compliance

To understand compliance it is best to first examine the choices made as a whole. Figure 3.1 shows the choices made by the Receivers as well as the theoretically optimal behavior. The size of the blue circles illustrates the number of participants who made the choice. The larger the circle the more participants chose a structure. The category shows the info structure and the message sent[13]. As a reference we reiterate the list of information structures from the previous section:

Information Structures:

A - Full information before period 1

B - No information

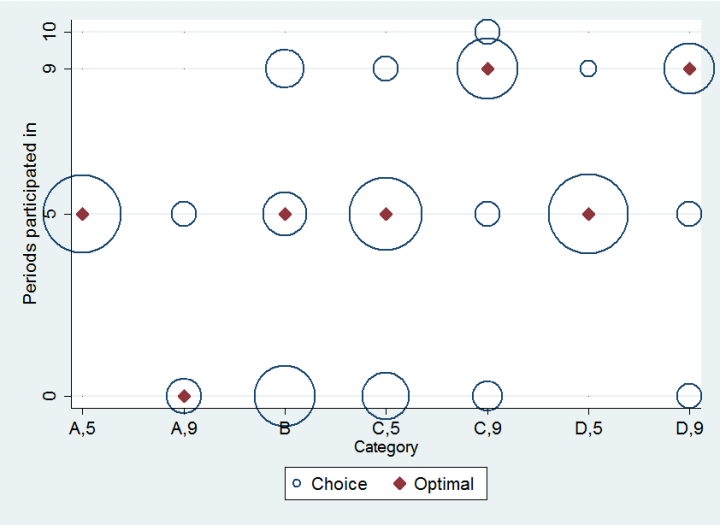
C - Partial information after 5 periods

D - Full information after 5 periods

Figure 3.1 makes clear that the compliance rates are quite different across the information structures. 100% of subjects optimally responded to receiving the message "The threshold period is 5." and 66.67% to "The threshold period is 9." under information structure A. Only 26.10% of subjects responded optimally to information structure B. 65.38% of subjects optimally responded to receiving the message "The threshold period is 5." and 63.16% to "The threshold period is 9." under information structure C. 95.24% of subjects optimally responded to receiving the message "The threshold period is 5." and 66.67%

[13] B has no message to the Receiver so is not split into two.

Figure 3.1: Optimal and chosen number of periods participated in

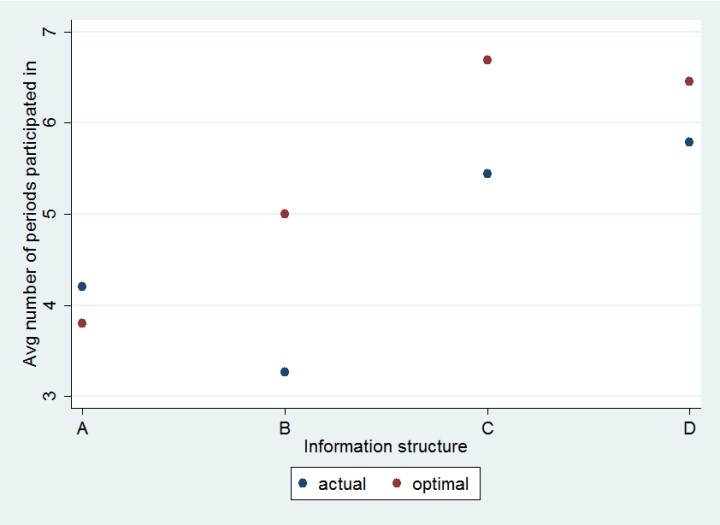


to "The threshold period is 9." under information structure D. Across all information structures and messages, 68.25% of the subjects responded optimally to the information they received. There is a direct relationship between the compliance rates and the informativeness of the information structure. A, which is the most informative, had the best compliance rate followed by D which is less informative due to the time delay. C had an even worse compliance rate and B had by far the worst. These aggregate rates are displayed later in table 3.3.

Given the compliance rates of the subjects, we now compare how well the different information structures perform compared to one another. Figure 3.2 shows the optimal average number of periods participated in and was weighted using the realized probabilities of each of the states and messages.

There is also a trend in the average periods participated in comparison to the optimal average. When presented with information structure A subjects actually on average participate in more periods than would be theoretically optimal. However, for the remaining three structures, the average actual number of periods participated in is lower than the optimal number, where the gap is strictly increasing in the amount of uncertainty contained in the information structure. Overall, D, i.e. providing full information after 5 periods,

Figure 3.2: Optimal and chosen average number of periods participated in



outperforms all other information structures. Note that this is true even though some subjects chose to participate in 10 periods after receiving information structure C. On the other hand upon getting C many subjects chose to stop immediately after 0 periods which is the driving force behind D outperforming C. We conclude that the knowledge that the information received after period 5 is accurate makes the subjects more willing to start participating. This speaks against the theoretical findings of section 3.2 which suggest C is optimal.

Let us now present a quantitative analysis of this qualitative result by checking whether the number of periods participated in for each information structure is significantly different from the optimal amount. Table 3.3 shows that just as illustrated in figure 3.2 the actual number of periods participated in is significantly lower than the optimal number of periods participated in for information structures B and C. A and D have actual means insignificantly different from the theoretical optimum. Both C and D outperform information structures A and B significantly, but as can be seen in the table the average periods participated in for D is insignificantly larger than that of C[14]. This is an important point to emphasize since this shows that the only real benefit for the Sender comes

[14]With a p-value of 0.255.

from delaying the information, and the addition of obfuscation does improve her outcome.

Table 3.3: Optimal and chosen average number of periods participated in by info structure

	Actual mean	Optimal mean	Diff	Compliance
Periods participated in for A	4.2 (25)	3.8	0.4 (0.2957)	92.00%
Periods participated in for B	3.26087 (23)	5	-1.73913** (0.0370)	26.10%
Periods participated in for C	5.35556 (45)	6.68889	-1.33333** (0.0125)	64.44%
Periods participated in for D	5.78788 (33)	6.45455	-0.66667 (0.1098)	84.85%

The numbers in parentheses under the means are numbers of observations. The number in parentheses under the difference is the p-value. Estimates are from one-sample t tests where we compared the actual mean to the optimal mean. Note that ** indicates $p < 0.05$.

To summarize, for the two structures with accurate information about the true threshold period, subjects did not participate in a number of periods that was statistically significantly different from the optimal number, while in the two structures with either no information or sometimes wrong information about the true threshold periods, subjects participated in statistically significantly fewer periods than optimal. There is hence an effect on the difference between average actual and actual optimal participation of the informativeness of the information structure. In addition, as mentioned earlier and as can be seen from the final column in table 3.3, informational content also has an effect on the compliance rate in that it is increasing in informativeness of the information structure. We will explore this finding in regards to risk attitudes shortly. First we finish the compliance data by showing the proportion of subjects who chose what information structure.

Table 3.4: Frequency of each information structure

Information Structure	Percentage	Number
A	22.22%	28
B	11.11%	14
C	34.13%	43
D	32.54%	41

Table 3.4 summarizes how frequently each information structure was chosen by par-

ticipants in the role of Senders, even though not all of these structures ended up being implemented. Information structures C and D were by far the most popular information structures to choose. These two structures theoretically predict the highest amount of participation in expectation under risk neutrality and it seems that subject Senders expected to achieve the highest participation by choosing either of these two information structures. This does not only show that the majority of subjects correctly understood our set-up but also that the theoretical results practically appealed to subjects.

Risk attitudes

As previously mentioned, the participation choices depicted in figure 3.1 seem to suggest that choices are driven by different risk attitudes of subjects, especially since the compliance is a function of information (and timing). The optimal response in the figure was calculated assuming risk neutrality. Let us extend our compliance analysis by allowing for different risk attitudes using the CARA utility realization[15]:

$$u(c) = \begin{cases} -\frac{1}{r}exp(-ry) \quad for \quad any \quad r \neq 0 \\ y \quad if \quad r = 0 \end{cases}$$

The parameter r indicates whether an agent is risk-neutral ($r = 0$), risk-averse ($r > 0$), or risk-loving ($r < 0$).

For information structure A, risk attitudes *do not* impact the optimal choices. For any value of r, the agent should participate in 5 periods if the threshold period is revealed to be 5 and should participate in 9 periods if the threshold period is revealed to be 9.

For information structure B, there are only two potential optimal choices: participating in 0 periods or participating in 5 periods. Participating in 0 periods is optimal if $r > 0.045$ and participating in 5 periods is optimal if $r \leq 0.045$. Note that going to 9 is never optimal because it leads to a guaranteed payoff of 9 which is smaller than a

[15]The functional form for CARA was adapted from Barseghyan et al. (2018).

guaranteed payoff of 10 from not starting to participate.

For information structure C, we have to start with the choice at period 5. If the message is that the threshold period is 5, then the agent should stop participating. But if the message is that the threshold period is 9, then risk attitude determines whether it is more optimal to stay at 5 or go to 9. It is optimal to stay at 5 for $r < -0.10$ and it is optimal to go to 9 if $r \geq -0.10$. To conclude the analysis of the optimal strategy, we have to see what the agent chooses at period 0 given what they would choose at period 5. For $r < -0.10$, the agent will optimally start and then stop at 5 no matter what the message. For $-0.10 \leq r < 0.474591$, the agent will start and stop at 5 after receiving the message that the threshold period is 5 and continue until 9 after receiving the message that the threshold period is 9. And for $r \geq 0.474591$, the agent will never start.

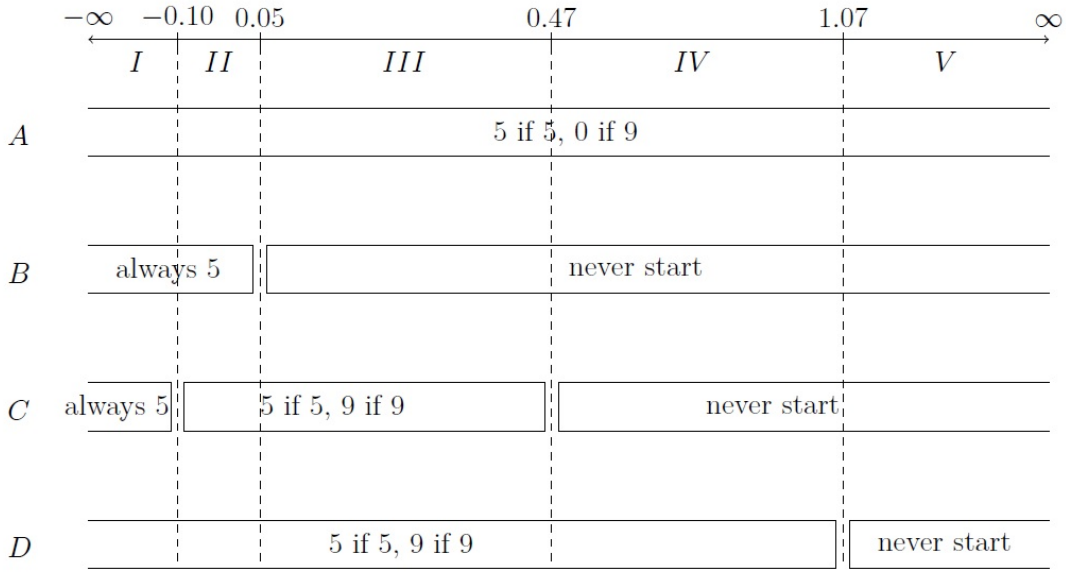
For information structure D, if the agent has participated until receiving the message, then she will always conform with the message. This leaves the decision whether she will start to participate or not. She will choose not to participate for $r > 1.07146$ and will choose to start participating for $r \leq 1.07146$.

We summarize the optimal participation choices for all values of r separately for each information structure in figure 3.3.

Let us denote the ranges from left to right as I (risk-loving), II (close to risk neutral), III (slightly risk-averse), IV (very risk-averse), and V (extremely risk-averse)[16]. For risk-category I, D induces the highest participation in expectation. This is because for the LEAD strategy, risk-loving agents are willing to take a gamble and stop participation after 5 periods if they get a message that the true threshold period is 9. For risk categories II and III, C - our approximate LEAD strategy - is expected to induce the highest amount of participation in expectation, for risk category IV, D is expected to induce the highest amount of participation in expectation. This is because agents faced with LEAD decide to never start participating. For risk category V, A is the only information structure that

[16]We have used the optimal responses to each information structure and message according to risk category II to calculate the optimal in figures 1 and 2 and in table 3.

Figure 3.3: Optimal participation choices for all values of r



induces in expectation a non-zero amount of participation and, hence, leads in expectation to the highest amount of participation. We reproduce figure 3.1 and figure 3.2 given the optimal choices for each risk-profile separately in figures 3.4 and 3.5 respectively.

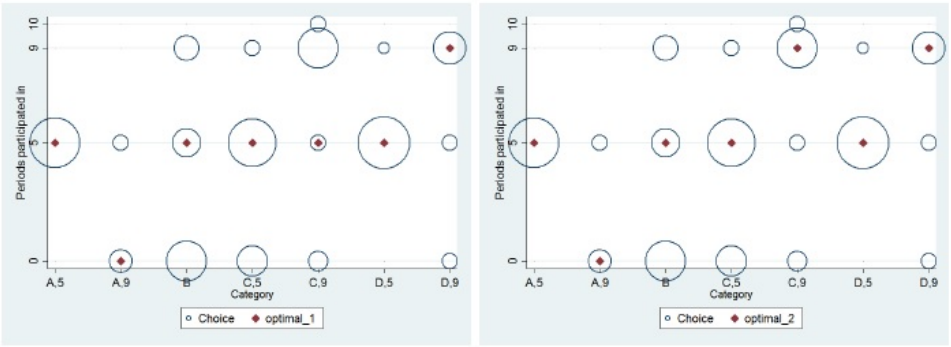
We can also re-derive the compliance rates assuming each different risk-attitude level. These are summarized in table 3.5.

Table 3.5: Compliance rate for each different risk attitude category

Information structure	I	II	III	IV	V
A	92%	92%	92%	92%	92%
B	26.10%	26.10%	52.17%	52.17%	52.17%
C	20.00%	64.44%	64.44%	22.22%	22.22%
D	84.85%	84.85%	84.85%	84.85%	6.06%
Total	60.32%	68.25%	73.02%	57.94%	37.30%

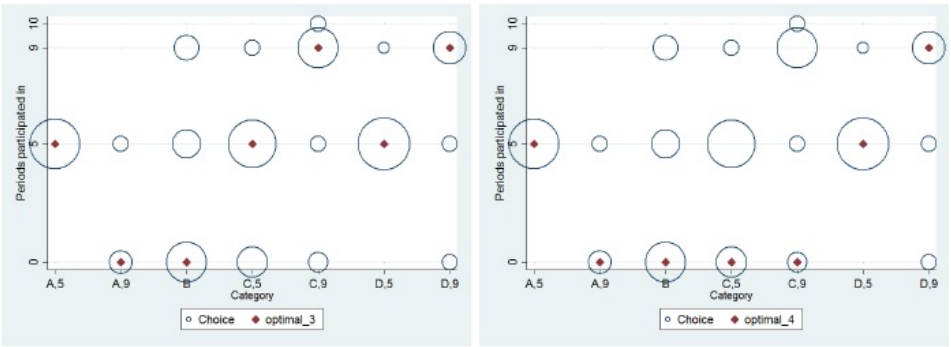
We can see that the largest overall compliance happens if we assume that the risk parameter r of subjects was laying in risk-attitude category III. Experimental findings have actually shown that real-life coefficients of absolute risk-aversion lie in this interval of r. Take as example Beetsma and Schotman (2001) who estimate the CARA parameter to be $r = 0.12$.

Figure 3.4: Optimal and chosen number of periods participated in by information choice and message, by risk level



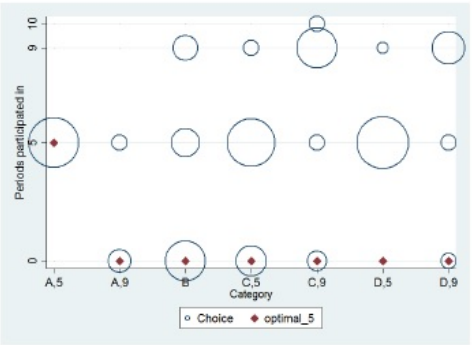
(a) I - risk-loving

(b) II - close to risk-neutral



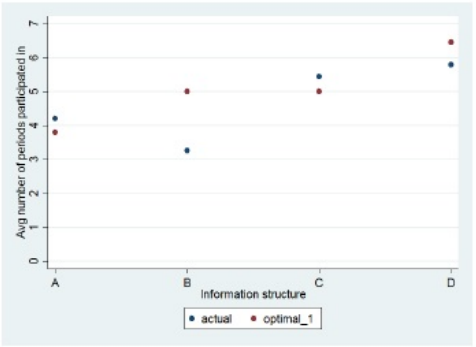
(c) III - slightly risk-averse

(d) IV - very risk-averse

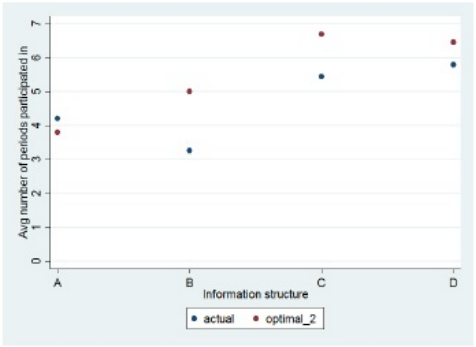


(e) V - extremely risk-averse

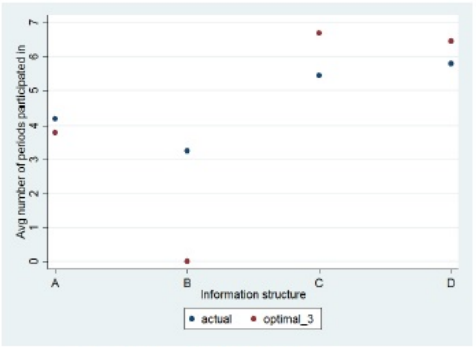
Figure 3.5: Optimal and actual average number of periods participated in by information structure, by risk level



(a) I - risk-loving

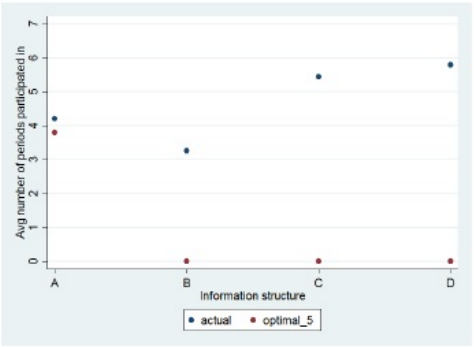


(b) II - close to risk-neutral



(c) III - slightly risk-averse

(d) IV - very risk-averse



(e) V - extremely risk-averse

The most interesting analysis happens however if we allow subjects to display varying degrees of risk-attitude, in which case we can explain 88.89% of all Receivers' decisions[17]. In addition, if we limit ourselves to these 88.89% and regions II, III, and IV and re-calculate the theoretical optimal of C by adding together the weighted optimals in each region, we get a theoretical average of 5.01 which is not statistically significantly different from what our subjects actually chose. In other words, allowing for a reasonable range of heterogeneity in risk attitudes makes information structure D theoretically optimal since it's optimum is the same across II, III, and IV. Heterogeneity in risk attitudes hence provides a reasonable explanation for the behavior which we find in our compliance section. In addition, this highlights a potentially important point about information design in that it fails to account for said heterogeneity, making it potentially not a good guide in practice.

3.4.2 Reciprocity

To isolate the reciprocity effect we compare the data in which the Sender's choice was implemented against the data in which the computer randomized the choice. Table 3.6 presents a summary of how the average number of periods participated in varies by the information structure the Receiver faced, the true threshold period, as well as by who chose the information structure. We can see that the average number of periods participated in is weakly higher for every information structure and threshold period whenever the information structure was chosen by a participant Sender compared to the computer. A possible explanation of the phenomenon is that the Sender is more present in the Receivers mind when the information structure was chosen by the Sender. This however does not seem like a very convincing explanation since during the practice rounds, each subject takes on the role of Sender and Receiver and also acts as the Sender during the first

[17]The choice of participating in 5 periods after receiving full information that the threshold period is 9 immediately or after 9 periods, participating in 9 periods when receiving no information or after receiving the message that the state is 5 from information structures C or D, or participating in 10 periods cannot be explained with risk-attitudes.

interaction of the payoff-relevant round. Clearly altruism can also play a role but as long as it is not too heterogeneous between the two groups it should be controlled for in the comparison.

Table 3.6: Average number of periods participated in by information structure, true threshold period, and who chose the information structure

Information structure	Threshold period	Sender	Computer
A	5	5.00 (9)	5.00 (10)
	9	1.22 (5)	0.00 (1)
B	5	4.50 (2)	3.23 (13)
	9	4.50 (2)	2.50 (6)
C	5	5.13 (16)	4.47 (15)
	9	6.67 (9)	6.40 (5)
D	5	5.40 (10)	5.00 (11)
	9	8.42 (7)	4.60 (5)

The number of observations in each cell is in parentheses.

This qualitative finding is confirmed by a quantitative analysis, the results of which are depicted in table 3.7. In the first specification we can see that overall, the number of periods participated in is significantly higher when the information structure was chosen by the Sender rather than the computer. Given the large difference in the occurrences of B in the second specification, we exclude info structure B in our comparison of average periods participated in. While the mean is still higher when the Sender chose the information structure than when the computer chose it, the difference is no longer statistically significantly different.

Based on our analysis in the previous section, we are most interested in the information structures C and D which had the surprising result that D outperforms C. The next two specifications compare the average number of periods participated in just for C and D. We can see that the number of periods participated in for C is not statistically significantly different depending on whether the information structure was chosen by the Sender or the computer but for D it is. In the last two specifications we can see that this difference is largely driven by the behavior of subjects when the state is 9. To illustrate the behavioral

Table 3.7: Difference in periods participated in based on whether the information structure was chosen by the computer or by a participant Sender

	Sender mean	Computer mean	Diff
Periods participated	5.4667	4.3030	1.1636**
	(60)	(66)	(0.0345)
Periods participated (excluding B)	5.535714	4.829787	0.7059271
	(56)	(47)	(0.2041)
Periods participated (only C)	5.68	4.95	0.73
	(25)	(20)	(0.4838)
Periods participated (only D)	6.647059	4.875	1.772059**
	(17)	(16)	(0.0273)
Periods participated (only D, state = 5)	5.4	5	0.4
	(10)	(11)	(0.3062)
Periods participated (only D, state = 9)	8.428571	4.6	3.828571*
	(7)	(5)	(0.0598)

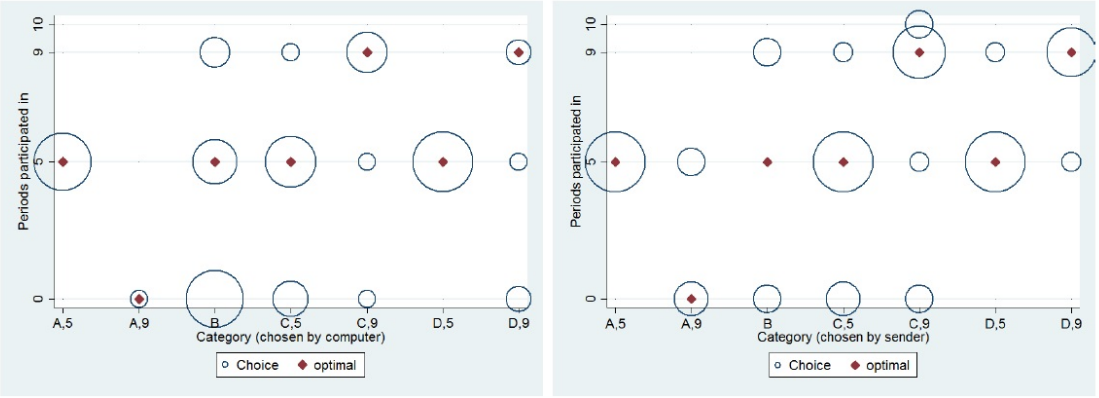
The numbers in parentheses under the means are numbers of observations. The number in parentheses under the difference is the p-value. Estimates are from two-sample t tests with unequal variance. Note that * indicates $p < 0.1$ and ** indicates $p < 0.05$.

difference, we have re-produced figure 3.1 separately for when the info structure was chosen by the computer and when it was chosen by the Sender in figure 3.6.

Due to the different realized distributions of structures and realizations, let us compare the difference between average actual and optimal number of periods participated in by the Receivers, separately for the computer and subject Senders. The results are depicted in figure 3.7, where the information structures are ordered by their "informativeness". We can see that when the Sender chooses the information structure, then the Receiver participates on average in more periods than optimal following information received according to structure A and the optimal amount following information received according to structure D. We can also surprisingly see that the gap when the info was sent by a subject Sender is smaller for structure B than for structure C.

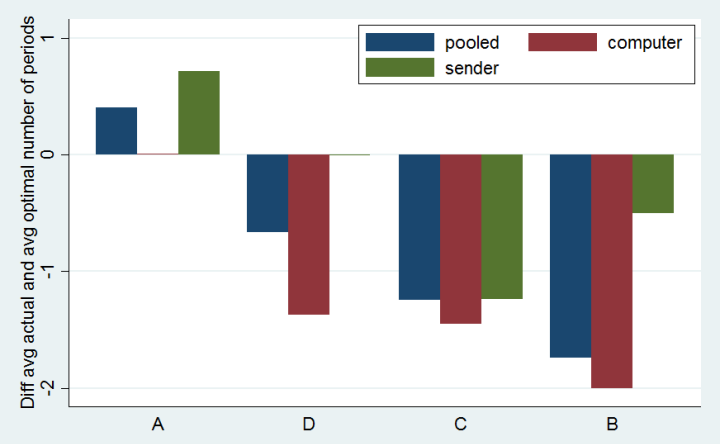
We can further see that when the computer chooses the information structure, then the Sender participates on average the optimal number of periods following information received according to structure A and fewer than the optimal number of periods follow-

Figure 3.6: Optimal and chosen number of periods participated in by information choice and message, by who chose the info structure



(a) Info structure chosen by computer (b) Info structure chosen by Sender

Figure 3.7: Optimal and chosen number of periods participated in by information choice



ing information received according to any other information structure. The difference decreases in the amount of information communicated to the Receiver. Let us interpret these findings in terms of the reciprocity definition by Charness (2004), referenced in footnote 9. Receivers exhibit positive reciprocity towards the Sender if the Sender chose information structure A. This effect is statistically significant at the 10% level with a one-sided alternative. Note that this finding is not robust to outliers. Consider figure 3.6. What drives the difference is that some subjects respond with participating in 5 periods to receiving the message "The true threshold period is 9." in information structure A. This behavior seems to be a mistake. Even if the subjects who made this choice had wanted to

benefit the Sender, they could have improved both their own and the Sender's payoff by continuing participation all the way to 9. Changing it to 9 would increase the Receiver's payoff by \$4 and the Sender's payoff by \$6. Eliminating these observations leads to perfect compliance with $A, 5$ and $A, 9$ no matter who sends the information structure. We hence conclude that there is no positive reciprocity in information structure A. We might want to interpret the behavior in response to information structure D as positive reciprocity since the difference is also significantly different at the 10% level. All the subjects are willing to participate in a positive number of periods. Moreover, in response to $D, 5$, some subjects are willing to participate in 9 rather than 5 periods, which increases the Sender's earnings by \$6 while it reduces own earnings by \$4. However, we might be hesitant to call this reciprocity since the average participation coincides with the theoretical optimum when the Sender chose the information structure. Hence, there is an attribution effect (the fact that the sender chose as opposed to a computer randomizing has an effect) but reciprocity does not seem to be the correct concept to describe it. For C, the difference is not statistically significant and for B, we have collected so few observations with a subject Sender that a conclusion is not possible.

3.5 Conclusion

Ely and Szydlowski (2020) recently made an interesting contribution to the Bayesian Persuasion literature by analyzing a principal-agent setup where the principal provides information, as opposed to monetary incentives, to extract effort from the agent. We have discretized their binary framework and studied it in the laboratory. Our results show that subjects do not deviate from the theoretically optimal responses when presented with full information before acting - which is equivalent to the typical Bayesian Persuasion setting - or delayed full information. However, subjects exert less effort than predicted when presented with no information or delayed partial information. The informativeness of

the principal's choice of information structure can then be seen as immediately affecting the willingness of agents to comply with the theoretical optimum, and in addition the informativeness also affects how far away the average participation is from the theoretical optimum. Additionally we find that there is no real additional benefit to the use of a signal rather than providing the agents with full information when using delayed information disclosure. This calls into question the optimality of Ely and Szydlowski's LEAD strategy in practice. We provide a plausible explanation for this in allowing for heterogeneity of risk attitudes among our subjects. In doing so we are able to explain 88.89% of their choices and we also find that by recalculating the theoretical optima with this heterogeneity in mind, the LEAD strategy is beaten by delayed full information both theoretically as well as in practice. This suggests and important avenue for further research in finding the theoretically optimal information disclosure policy under heterogeneous risk attitudes.

While we expected to see a similar response to that of the subjects in Charness (2004) in that for the better information structures the response would be positive reciprocity and negative for the worse information structures, we find no such result. There seems to be a clear attribution effect for information structure D in which the true state of the world is revealed after 5 periods, but it does not fall in line with the behavior discussed by Charness which would see the subjects playing the theoretical optimum against the computer. We observe the subjects playing the theoretical optimum when another subject sent the information and much less than is optimal when faced with the computer. It should also be mentioned that in the wage rate setting of Charness it is much clearer what constitutes "nice" and "bad" behavior on the part of the principal which is setting a higher and lower wage rate respectively. In our setting it is perhaps not immediately clear what would be the representative "nice" and "bad" behavior, which would explain why we do not observe the reciprocity and punishment we expected. The main take-away from this is that in real-life settings where the principal only provides information as an incentive, the principal can act closely to the optimal without punishment concerns.

3.6 Appendix

3.6.1 Screenshots

The following figures show the relevant screenshots from our experiment. Figure 3.8 shows the Sender's decision screen during the information stage of the interaction. Figure 3.9 shows the first screen of the participation stage on which the Receiver is informed about according to which information structure she will receive information about the threshold period and about whether this information structure was chosen by the Sender or by the computer. Figures 3.10 and 3.12 show the Receiver's decision screen before and after the Receiver has received a message. Figure 3.11 gives an example of a message screen.

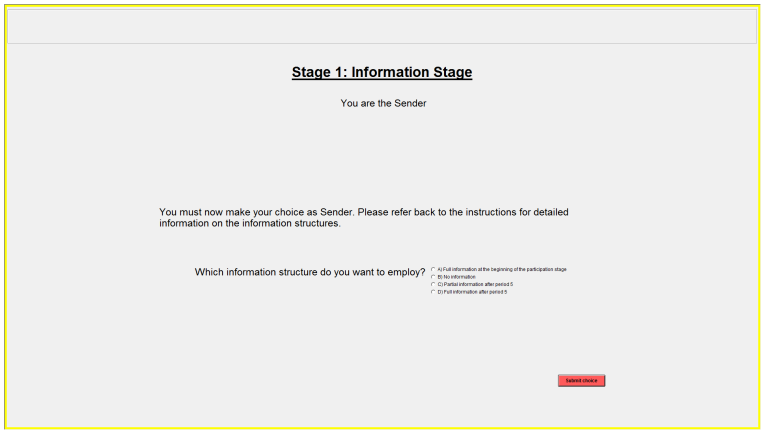


Figure 3.8: Sender's decision screen

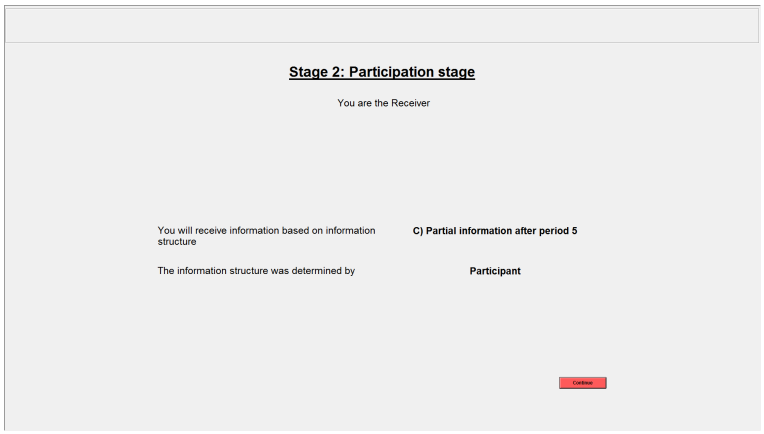


Figure 3.9: First screen of participation stage

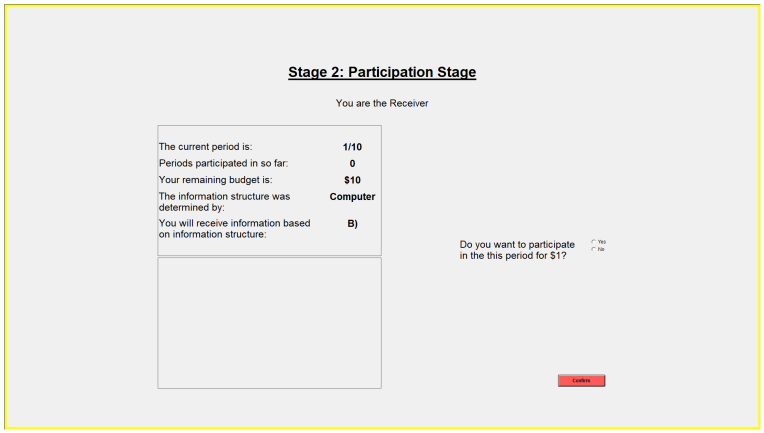


Figure 3.10: Receiver's decision screen *before* she receives a message

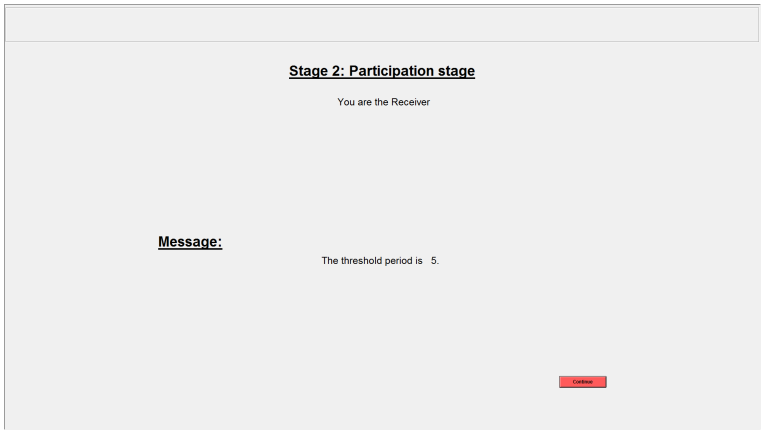


Figure 3.11: Message screen

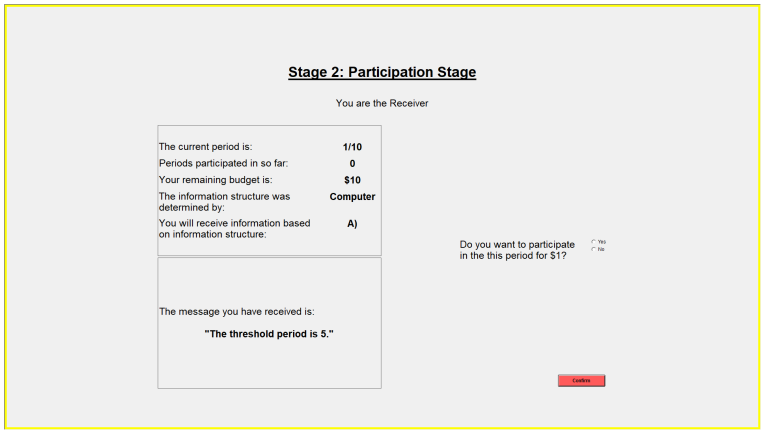


Figure 3.12: Receiver's decision screen *after* she receives a message

3.6.2 Experimental instructions

In this section of the appendix, we reproduce the experimental instructions. A copy of these was handed out to the participants *after* they signed a consent form but before the experiment started. These instructions were read out aloud.

Overview

This experiment consists of interactions between an agent who sends information (a Sender) and an agent who receives and acts on the information (a Receiver).

The act in question is that the Receiver must choose how many consecutive periods to participate in. To this end, in each period, the Receiver can choose to continue participating or end the interaction. Choosing to participate in a period incurs a cost, which is paid for using a budget that the Receiver is endowed with. The Receiver will win a reward if he or she participates for sufficiently many periods before ending the interaction. How many periods are sufficient to win the reward is randomly determined before the start of the interaction, and it can be either 5 with probability 2/3 or 9 with probability 1/3. We will refer to the number of the last period a Receiver must participate in to win the reward as the threshold period, so if the Receiver will win the reward after participating in 5 periods, we call the threshold period 5, and if the Receiver will win the reward after participating in 9 periods, we call it 9. We will call this part of the interaction the Participation Stage.

The information that the Sender can send to the Receiver is regarding this unknown threshold period. The Sender instructs the computer on when and how much information about the threshold period is going to be sent to the Receiver. We call this part of the interaction the Information Stage, and it precedes the Participation stage.

An interaction then goes as follows. First the threshold period is randomly determined but not revealed to either of the two agents. Then follows the Information Stage, and then the Participation Stage.

You will take part in two interactions. In the first you will take on the role of the

Sender for another participant in this room who takes the role of the Receiver. In the second you will take on the role of the Receiver and will be paired with another participant who takes on the role of the Sender. You are guaranteed not to face the same participant in both of these interactions and both interactions are anonymous.

We now explain each of the stages in detail and then explain how the earnings are calculated.

Stage 1: Information stage (You are the Sender)

In this stage, the Sender instructs the computer on when and how to send information to the Receiver about the threshold period. There are four possible information structures, i.e. ways in which the computer communicates information about the threshold period to the Receiver, the Sender can choose from:

A) *Full information at the beginning of the participation stage*:

The computer will send the message "The threshold period is 5." if the threshold is drawn to be the 5th period and will send the message "The threshold period is 9." if the threshold is drawn to be the 9th period. The message is sent right at the beginning of the participation stage.

B) *No information*:

The computer never sends a message with information about the true threshold period.

C) *Partial information after period 5*:

The computer can send the message "The threshold period is 5." or "The threshold period is 9." If the computer sends the message "The threshold period is 5.", then the true threshold period is 5. But if the computer sends the message "The threshold period is 9.", then the true threshold period is either 5 with probability 2/5 or

is 9 with probability 3/5. The message is sent after period 5 if the Receiver has participated until then.

D) *Full information after period 5*:

The computer will send the message "The threshold period is 5." if the threshold is drawn to be the 5th period and will send the message "The threshold period is 9." if the threshold is drawn to be the 9th period. The message is sent after period 5 if the Receiver has participated until then.

After the Sender has made his or her decision, with probability 1/2 the Sender's choice is implemented and, with probability 1/2, one of the four information structures is randomly chosen and implemented by the computer. Which information structure is implemented as well as whether it was chosen by the Sender (another participant), or randomly by the computer will be reported to the Receiver right at the beginning of the second stage. Note that the choice of information structure does not directly impact the earnings of the Sender, which instead only depends on the participation choices made by the Receiver that the Sender is matched with. Further detail will be provided below.

Stage 2: Participation stage (You are the Receiver)

In this stage the Receiver will receive and then act on the information about the threshold period. On the first screen of the stage, the Receiver will be informed about which information structure will be used and whether it was chosen by the computer or by the Sender (another participant).

If the information structure chosen is A (i.e. the Receiver learns about the threshold period before having to make any decisions) then the second screen of the stage shows a message containing the threshold period. If the chosen information structure is B, C, or D, there is no message screen displayed at this time.

Now it is time for the Receiver to act, and the next screen is the period 1 participation screen on which the Receiver must choose whether to participate in this period or not.

This is the first of 10 identical such screens. The decision to participate in the period is made by choosing "yes" or "no" on the menu and confirming the choice. As mentioned previously, participating in a period is costly in that choosing "yes" deducts $1 from the budget of $10 that the Receiver is endowed with. For your reference, your current budget is displayed on the top left of the screen together with the current period, the number of periods you have participated in, the information structure, and whether the information structure was chosen by the Sender or by the computer. Choosing "yes" further leads to another identical participation screen with updated values. The Receiver is then faced with the same participation decision, and this continues until either the Receiver runs out of money or he or she decides to not participate in a period by choosing "no" on a participation screen.

If the information structure chosen was C (i.e. the Receiver gets partial information about the threshold period after 5 periods) or D (i.e. the Receiver learns about the threshold period after 5 periods), a message screen will appear with the relevant message after choosing "yes" on the 5th period participation screen and confirming the choice. After acknowledging the message the 6th period participation screen follows.

If the Receiver has participated in at least as many periods as the threshold period (i.e. has chosen "yes" on the 5th period participation screen in case the threshold period is 5 or has chosen "yes" on the 9th period participation screen in case the threshold period is 9) when the stage ends, the Receiver earns a reward of $8. Whatever amount of the budget is not spent on participation will be part of the Receiver's earnings. We now explain in more detail how the earnings for both the Sender and the Receiver are calculated.

Earnings calculation

The earnings are calculated the same way in each of the two interactions. Only one of them is chosen randomly with equal probability to be paid out along with the $5 show-up fee.

The Receiver's earnings if he or she participates at least until the threshold period are:

$$Earnings\ for\ Receiver\ with\ reward = \$10 - \$1 * number\ of\ periods\ participated\ in + \$8.$$

The Receiver's earnings if he or she does *not* participate until the threshold period are:

$$Earnings\ for\ Receiver\ without\ reward = \$10 - \$1 * number\ of\ periods\ participated\ in.$$

The Sender's earnings regardless of whether the Receiver earns the reward or not are:

$$Earnings\ for\ Sender = \$1.5 * number\ of\ periods\ the\ Receiver\ participated\ in.$$

All possible earnings combinations are summarized in Table 1 below.

Table 1: Earnings

Threshold	Role	Number of periods Receiver participated in										
		0	1	2	3	4	5	6	7	8	9	10
5	Receiver	\$10	\$9	\$8	\$7	\$6	\$13	\$12	\$11	\$10	\$9	\$8
9	Receiver	\$10	\$9	\$8	\$7	\$6	\$5	\$4	\$3	\$2	\$9	\$8
5 or 9	Sender	\$0	\$1.50	\$3	\$4.50	\$6	\$7.50	\$9	\$10.50	\$12	\$13.50	\$15

Understanding questions and practice rounds

Now you must answer correctly a series of understanding questions about the experiment. Then you will take part in 5 practice rounds. Each practice round consists of two interactions. Just as in the payoff-relevant round, you will first make a choice as a Sender and then as a Receiver. However, during these practice rounds, you will *not* be paired with other participants. This means that during the participation stage if the info structure implemented was chosen by a Sender, it is the one you chose as Sender during the information stage. The practice rounds are meant for you to familiarize yourself with the screens and tasks of both roles. We strongly urge you to take these rounds seriously as

this might help you increase your earnings in the payoff relevant round that will follow the practice rounds. However, all the choices that you make in the practice rounds are unpaid and they do not affect the actual experiment.

Only after the understanding questions and the 5 practice rounds do the two payoff relevant interactions start, where you are paired with two actual participants in the room. The start of the payoff-relevant interactions will be clearly marked in the program that is used for the experiment.

Final Summary

Before we start, let us remind you of the following:

1. The threshold period (which is the period the Receiver must participate up to and including to win the reward) is either 5 with probability 2/3 or 9 with probability 1/3.

2. In the Information stage the Sender chooses from the following 4 information structures:

   A) The threshold period is revealed before period 1.

   B) The threshold period is never revealed.

   C) More information about the threshold period is given after period 5.

   D) The threshold period is revealed after period 5.

3. With 1/2 probability the Sender's choice is used, and with 1/2 probability the computer randomly chooses the information structure.

4. In the Participation stage the Receiver chooses to participate until he or she wants to stop, and earns a reward of $8 if he or she has participated up to and including the threshold period.

5. Participation costs $1 per period and is paid for using a budget of $10.

6. The Sender earns $1.50 per period the Receiver has participated in.

7. Each participant in the experiment will be a Sender in one interaction and a Receiver in another interaction - you do not interact with the same person twice.

8. The total earnings from the experiment are the earnings in one of the two interactions, randomly determined, plus $5 show-up fee.

BIBLIOGRAPHY

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Ai, C. and Chen, X. (2007). Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141(1):5–43.

Antoine, B. and Dovonon, P. (2018). Robust estimation with exponentially tilted hellinger distance. *Working Paper*.

Au, P. H. and Li, K. K. (2018). Bayesian persuasion and reciprocity: theory and experiment. *Available at SSRN 3191203*.

Aumann, R. J., Maschler, M., and Stearns, R. E. (1995). *Repeated games with incomplete information*. MIT press.

Barseghyan, L., Molinari, F., O'Donoghue, T., and Teitelbaum, J. C. (2018). Estimating risk preferences in the field. *Journal of Economic Literature*, 56(2):501–64.

Beetsma, R. M. and Schotman, P. C. (2001). Measuring risk attitudes in a natural experiment: data from the television game show lingo. *The Economic Journal*, 111(474):821–848.

Charness, G. (2004). Attribution and reciprocity in an experimental labor market. *Journal of labor Economics*, 22(3):665–688.

Charness, G., Gneezy, U., and Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, 149(C):74–87.

Charness, G. and Levin, D. (2005). When optimal choices feel wrong: A laboratory study of bayesian updating, complexity, and affect. *American Economic Review*, 95(4):1300–1309.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Chaudhuri, S. and Renault, E. (2017). Score tests in GMM: Why use implied probabilities? *JPE Forthcoming*.

Chen, S. and Schildberg-Hörisch, H. (2018). Looking at the bright side: The motivation value of overconfidence. *IZA Discussion Paper*.

Chen, X., Hong, H., and Shum, M. (2007). Nonparametric likelihood ratio model selection tests between parametric likelihood and moment condition models. *Journal of Econometrics*, 141(1):109 – 140. Semiparametric methods in econometrics.

Corcoran, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, pages 967–972.

Csiszár, I. (1995). Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, 68(1-2):161–186.

Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565.

DellaVigna, S. and Pope, D. (2018). What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069.

Ely, J. C. and Szydlowski, M. (2020). Moving the goalposts. *Journal of Political Economy*, 128(2):000–000.

Eriksson, T., Poulsen, A., and Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, 16(6):679–688.

Fehr, E., Kirchler, E., Weichbold, A., and Gächter, S. (1998). When social norms overpower competition: Gift exchange in experimental labor markets. *Journal of Labor economics*, 16(2):324–351.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178.

Fréchette, G. R., Lizzeri, A., and Perego, J. (2019). Rules and commitment in communication. *CEPR Discussion Paper No. DP14085*.

Hall, P. and Horowitz, J. L. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica: Journal of the Econometric Society*, pages 891–916.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Hansen, L. P., Hansen, P. G., and Mykland, P. A. (2016). Measuring belief distortions. *Unpublished Manuscript*.

Imbens, G., Johnson, P., and Spady, R. H. (1995). Information theoretic approaches to inference in moment condition models.

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

Newey, W. K. (2004). Efficient semiparametric estimation via moment restrictions. *Econometrica*, 72(6):1877–1897.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.

Nguyen, Q. (2017). Bayesian persuasion: evidence from the laboratory. *Work. Pap., Utah State Univ., Logan*.

Otsu, T. (2007). Penalized empirical likelihood estimation of semiparametric models. *Journal of Multivariate Analysis*, 98(10):1923–1954.

Ragusa, G. (2011). Minimum divergence, generalized empirical likelihoods, and higher order expansions. *Econometric Reviews*, 30(4):406–456.

Renault, E. and Wahlstrom, O. (2020). Probabilities implied by misspeci
ed moment conditions. *Unpublished Manuscript*.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of econometrics*, 2:881–935.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, 35(2):634–672.

Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.