Bounded Rationality and Noise : Theory, Methods and Experiments

Fabien Perez

INSTITUT POLYTECHNIQUE DE PARIS

ENSAE

IP PARIS

NNT : 2021IPPAG008

Thèse de doctorat

Bounded Rationality and Noise : Theory, Methods and Experiments

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'Ecole Nationale de la Stastique et de l'Administration Economique

École doctorale n°626 de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Sciences économiques

Thèse présentée et soutenue à Palaiseau, le 20 décembre 2021, par

Fabien Perez

Composition du Jury :

Yukio Koriyama Professeur, Ecole Polytechnique	Président
Marie Claire Villeval Directrice de Recherche, GATE-Université de Lyon	Rapporteur
Georg Weizsäcker Professeur, Humboldt-Universität zu Berlin (HU)	Rapporteur
Philippe Choné Professeur, Ensae Paris	Examinateur
Nagore Iriberri Professeur, University of the Basque Country	Examinatrice
Guillaume Hollard Directeur de Recherche, Ecole Polytechnique	Directeur de thèse

Remerciements – *Acknowledgments*

Je voudrais tout d'abord remercier mon directeur de thèse Guillaume Hollard pour son accompagnement tout au long de cette thèse. Je me souviendrais toujours de notre première rencontre dans son ancien bureau au département d'économie de l'Ecole polytechnique. Depuis ce jour, il m'a toujours soutenu dans les hauts et dans les bas. Nos longues discussions sur la recherche et sur la vie en général ainsi que les aventures que nous avons vécues ensemble m'ont beaucoup fait évoluer. Je lui en suis très reconnaissant.

I thank Marie Claire Villeval and Georg Weizsäcker for accepting to review my Ph.D. thesis, Philippe Choné, Yukio Koriyama, and Nagore Iriberri for accepting to be members of the jury. Our discussions through my Ph.D. and their research have truly inspired me, and it is a great honor to have them as reviewers and examiners of my dissertation.

J'aimerais remercier mes co-auteurs Radu Vranceanu, Héloïse Clolery et Inès Picard pour leur aide dans divers projets.

I would like to thank all CREST researchers for their help, through discussions or remarks. In particular, I am grateful to Philippe Choné and Yukio Koriyama – who were in my follow-up committee – and also Pierre Boyer, Xavier d'Haultfœuille, Bertrand Garbinti, Francis Kramarz, Yves Le Yaouanq, Laurent Linnemer, Matías Núñez, Vianney Perchet, Ivaylo Petev, Benoît Schmutz, Thibaud Vergé et Michael Visser.

J'aimerais également remercier les doctorants du CREST dont certains sont devenus des amis. Merci à Rémi Avignon et ses avis toujours tranchés, à Gwen-Jiro Clochard, mon partenaire de *behavioral* et mon binôme en charnière centrale, à Etienne Guigue et ses anecdotes du week-end, à Yannick Guyonvarch et son désamour des fonctions caractéristiques, à Béréngère Patault et sa bonne humeur, et à Elia Pérennès pour son écoute et ses conseils. Un grand merci à Lucas Girard pour toutes nos discussions philosophiques et méthodologiques, j'en ai appris beaucoup grâce à lui ! J'aimerais remercier aussi tous les autres doctorants que j'ai croisés et avec qui j'ai échangé – la liste est longue et j'en oublie sûrement – Reda Aboutajdine, Antoine Bertheau, Guidogiorgio Bodrato, Léa Bou Sleiman, Marion Brouard, Pauline Carry, Geoffrey Chinot, Pierre-

Edouard Collignon, Maddalena Conte, Thomas Delemotte, Léa Dubreuil, Antoine Ferey, Christophe Gaillac, Germain Gautier, Anaïs Goburdhun, Aymeric Guidoux, Morgane Guignard, Jérémy Hervelin, Morgane Hoffmann, Myriam Kassoul, Margarita Kirneva, Raphaël Lafrogne-Joussier, Alice Lapeyre, Alexis Larousse, Jérémy L'Hour, Hadrien Le Mer, Clémence Lenoir, Claire Leroy, Pauline Leveneur, Alexis Louaas, Esther Mbih, Denys Medee, Alfonso Montes-Sanchez, Martin Mugnier, Sandra Nevoux, Elio Nimier-David, Ivan Ouss, Louis-Daniel Pape, Felix Pasquier, Anasuya Raj, Emilie Sartre, Nicolas Schreuder, Yuanzhe Tang, Clémence Tricaud, Benjamin Walter, Ao Wang.

J'aimerais remercier Sri Srikandan et Waël Bousselmi pour leur aide sur les expériences au laboratoire du CREST. Je voudrais également remercier Teddy Arrif, Murielle Jules, Weronika Leduc, Lyza Racon, Arnaud Richet, Béatrice Tartrou, Fanda Traoré et Edith Verger pour leur aide dans les aspects logistiques et administratifs.

Je souhaiterais à présent remercier mes collègues de l'Ensae Paris, en particulier les coordinateurs d'enseignements avec lesquels j'ai beaucoup échangé, en particulier Morgane Cure et Jérôme Trinh, les assitants/coordinateurs en économie, Lucie Neirac, Wissal Sabbagh, Arthur Thomas mais aussi la direction des études dans son ensemble : Nicolas Apoteker, Laurent Davezies, Oumi Djoumoi, Anne Grillot, Guillaume Hu, Christine Imbert-Courtois, Bertille Jarry, Frédéric Loss, Audrey Mallet, Karine Pain, Rosalinda Solotareff. Merci égaelement à – la liste est longue et j'en oublie encore sûrement mais ils se reconnaîtront– Wasfe Achafar, Elisabeth Andreoletti-cheng, Michèle Arger, Charly Awad, Souleymane Bah, Marie-Christine Baker, Luciana Barbera, Pierre Biscourp, Nathalie Branger, Emmanuel Cavadini, Carole Delamare, Djamila Gherabi, Laurent Herco, Marylène Jubertie, Soufiane Kjidaa Christophe Lagarde, Jean-Luc Mérat, Françoise Moreels, Tarik Mouhtadi, Agnès Pelage, Philippe Pinczon du Sel, Thierry Raussin, Laurent Salmon, Marie-Françoise Schmitt, Patrick Tartrou, Sarah Teriitaumihau, Olivier Trouvé, Zied Zekri et aux services informatiques et logistiques dans leur ensemble. J'aimerais adresser un remerciement particulier à Anna pour ses accueils chaleureux et les services qu'elle rend à l'Ensae.

Je voudrais remercier mes plus proches amis à Paris qui m'ont également soutenu ces dernières années : Hugo, Olivier, Pierre, Tom, Tristan, Vincent, et évidemment mes colocataires Thibaut et Lucas. Je remercie aussi la bande de l'Insee, Odile, Julien et Paul-Armand.

Jules Depersin aurait pu apparaître déjà trois fois dans ces remerciements, dans la liste des doctorants du CREST, dans celle des membres de la direction des études mais aussi dans celle de mes plus proches amis à Paris. Pendant cette thèse, il aura aussi été mon co-bureau et mon partenaire de salle de sport. Merci à lui pour tous ces moments, toutes ces discussions qui m'ont fait évoluer sur tant de sujets différents !

J'ai une pensée particulière pour mes amis de la banlieue toulousaine, ceux qui

m'accompagnent depuis une quinzaine d'années, en particulier, Léa, Jérémi, Grégoire, Romain et Camille. Merci pour cela.

Je souhaiterais remercier Sophie pour son soutien indéfectible, son enthousiasme, son humour et ses attentions qui m'ont beaucoup aidé pendant cette fin de thèse.

Enfin, je tiens à remercier ma famille, en particulier mes parents Manuel et Maria, mon frère Thomas, ma sœur Marie, qui n'ont jamais cessé de me soutenir et de croire en moi, pendant les hauts et les bas de cette thèse, et plus généralement tout au long de ma vie.

Summary

Economists aim at understanding how agents make decisions in economic situations. The standard approach in the economic literature is to assume that agents are perfectly rational. However, the perfect rationality assumption is highly restrictive and does not perfectly capture how agents behave. Studying how bounded is the rationality of economic agents is thus a matter of prime importance. The present dissertation contributes to this research question in behavioral and experimental economics with a particular focus on behaviors in strategic environments and on noisy (rather than purely deterministic) decisions. The dissertation is composed of four chapters. Chapters 1 and 2 address behaviors in strategic interactions with different angles. Chapters 3 and 4 focus on the issue of noise in experimental tasks. Chapter 1 offers insight into how self-selection could promote rationality in strategic environments. Self-selection has seldom been studied under the scope of bounded rationality. In most experiments, once seated in the lab, subjects are forced to participate in all the tasks they faced. On the contrary, in many economic situations in the field (e.g. auctions, financial markets) agents can choose whether they want to participate. In this regard, understanding the consequences of self-selection is valuable. We assess the evolution of strategies when a self-selection stage is added in experimental games and study the potential drivers of self-selection into games. Chapter 2 proposes and tests a new model in behavioral game theory. The model assumes the existence of two types of players: confused players who do not form any beliefs and thus play randomly or naively, and strategic players who best respond to noisy beliefs regarding the probability of facing a confused or a strategic player. The model has interesting theoretical properties and fits experimental data with a single parameter. Chapter 3 calibrates measurement error in four risk-aversion tasks using a large set of existing test-retest experiments. Since measurement error can have a dramatic impact on statistical analysis, quantifying the noise in experimental measures is of particular importance. We also discuss the consequences of measurement error coupled with discrete approximations and small samples when performing linear regressions. Chapter 4 focuses on tasks aiming at eliciting trust. Using a test-retest experiment, we calibrate measurement error in standard trust measures, namely the behavior in a trust game and survey questions. We then discuss how cognitive skills and attention can drive noise in experimental tasks.

Résumé

La science économique s'intéresse à la façon dont les agents prennent des décisions dans des situations économiques. L'approche standard dans la littérature économique est de supposer que les agents sont parfaitement rationnels. Cependant, l'hypothèse de rationalité parfaite est très restrictive et peine parfois à expliquer le comportement empirique des agents. L'étude de la rationalité limitée des agents est donc une question centrale en économie. Cette thèse contribue à cette question de recherche en économie comportementale et expérimentale, en s'intéressant particulièrement aux comportements dans les environnements stratégiques ainsi qu'aux décisions bruitées (plutôt que purement déterministes). Ce manuscrit est composé de quatre chapitres. Les chapitres 1 et 2 abordent les comportements dans les interactions stratégiques sous différents angles. Les chapitres 3 et 4 s'intéressent à la question du bruit dans les tâches expérimentales.

Chapitre 1

Le chapitre 1 examine un mécanisme qui pourrait favoriser la rationalité dans les environnements stratégiques : l'auto-sélection. Il se base sur une observation simple. Dans la plupart des expériences, une fois assis dans le laboratoire, les sujets sont forcés de participer à toutes les tâches auxquelles ils sont confrontés. Au contraire, dans de nombreuses situations économiques (par exemple, les marchés financiers et les enchères), les agents peuvent choisir s'ils veulent participer. Cette étape d'auto-sélection qui se produit dans la vie économique peut influencer les comportements dans les situations économiques en question. Dans ce chapitre, nous concevons une expérience avec une étape d'auto-sélection explicite et évaluons l'évolution des stratégies des agents. Nous étudions en outre les moteurs potentiels de l'auto-sélection dans les jeux. Nous complétons notre analyse en examinant les preuves existantes concernant l'auto-sélection dans les environnements stratégiques. Enfin, nous discutons de la manière dont l'auto-sélection pourrait améliorer la validité externe des expériences de laboratoire.

Résultats: Les stratégies évoluent de manière significative lorsque l'on rajoute une étape d'auto-sélection. L'auto-sélection agit comme un filtre de l'irrationalité. Par exemple, moins de stratégies dominées sont jouées, la proportion de joueurs choisissant

les stratégies de l'équilibre de Nash augmente, et la sophistication stratégique des sujets est plus élevée. Cependant, les joueurs ne s'adaptent pas beaucoup aux changements dans le pool de sujets résultant de l'auto-sélection. En ce qui concerne les moteurs de l'auto-sélection, nous constatons que les joueurs averses au risque ont tendance à moins entrer dans les environnements stratégiques. En outre, une mesure de confiance que nous avons introduite, et qui est nouvelle dans le contexte des jeux expérimentaux, est corrélée avec la décision d'entrée. De plus, cette mesure de confiance est spécifique au jeu et reflète bien le sentiment de faire la bonne chose dans un jeu puisqu'elle est bien corrélée avec les stratégies jouées dans le jeu en question. Notre analyse sur l'effet de l'auto-sélection dans des situations hors du laboratoire conduit à des résultats similaires : les agents économiques auto-sélectionnés sont moins averses au risque, plus rationnels et plus homogènes. Nous espérons que cette étude suscitera un intérêt pour le mécanisme d'auto-sélection qui pourrait être mis en œuvre dans de nombreux contextes expérimentaux.

Chapitre 2

Le chapitre 2 introduit un nouveau modèle comportemental en théorie des jeux. Ce modèle rend compte de deux faits empiriques robustes : l'existence de sujets non stratégiques et l'hétérogénéité des croyances. En effet, plusieurs articles ont identifié, avec des méthodologies différentes, une proportion significative (entre 30% et 50%) de sujets confus qui ne sont pas capables de former des croyances. En ce qui concerne l'hétérogénéité des croyances, des recherches montrent que les sujets déclarent des croyances très variées. Le modèle que nous introduisons comporte deux types de joueurs : les joueurs incapables de former des croyances (not-forming beliefs ou encore NFB players) et les joueurs stratégiques. Comme leur dénomination le suggère, les joueurs NFB ne peuvent former aucune croyance et jouent de manière aléatoire ou naïve. Au contraire, les joueurs stratégiques forment des croyances en ce qui concerne la probabilité d'affronter un joueur NFB ou un joueur stratégique aussi sophistiqué qu'eux. Néanmoins, les joueurs stratégiques ont des croyances bruitées concernant la proportion de joueurs NFB et de joueurs stratégiques. Les joueurs stratégiques répondent au mieux à leurs croyances mais ignorent que les autres joueurs stratégiques peuvent avoir des croyances différentes.

Résultats: Nous introduisons le "better-beliefs model". Nous montrons l'existence d'une version à un paramètre du modèle dans laquelle les agents stratégiques ont des croyances hétérogènes concernant les adversaires qu'ils affrontent, grâce à des arguments de géométrie semi-algébrique. Le modèle a des propriétés théoriques intéressantes. En

particulier, lorsque le paramètre est à son minimum (respectivement son maximum), la distribution des stratégies prédites par le modèle est uniformément aléatoire (respectivement la stratégie d'un équilibre de Nash). Le paramètre représente à quel point un pool de sujets est stratégique. Enfin, le modèle explique et prédit correctement les données issues de deux précédentes études expérimentales.

Chapitre 3

Le chapitre 3 s'intéresse au problème des erreurs de mesure dans des tâches expérimentales classiques : les mesures d'aversion au risque. L'étude de l'aversion au risque est primordiale puisque de nombreuses décisions économiques peuvent être liées à l'aversion au risque relative des individus : investissement immobilier, assurance maladie, assurance vie, choix de carrière. Par conséquent, comprendre à quel point les agents économiques sont averses au risque peut être crucial pour prédire ou expliquer certains comportements économiques. Des tâches expérimentales d'aversion au risque, qui reposent majoritairement sur des choix entre différentes loteries, ont donc été conçues. Dans ce chapitre, nous essayons de calibrer la part de bruit (d'erreur de mesure) incluse dans ces tâches car, comme nous le verrons, les erreurs de mesure peuvent avoir des conséquences statistiques dramatiques. Nous calibrons d'abord la part d'erreur de mesure en utilisant cinq études différentes avec quatre tâches différentes et 16 ensembles de données différents. Nous effectuons des simulations pour évaluer l'impact de l'erreur de mesure couplée à un petit échantillon sur la signification des coefficients. Nous essayons également de déterminer si le fait que la plupart des mesures d'aversion au risque soient discrètes et non continues peut être problématique lors de la recherche de corrélations entre les variables.

Résultats: Nous trouvons que les mesures d'aversion au risque incluent une part significative de bruit : entre 35% et 60 % de la variance de la mesure est liée aux erreurs de mesure. Comme la théorie le prédit, cela mène à un biais d'atténuation dont le facteur est égal à environ un demi : les estimateurs des coefficients dans des régressions linéaires simples sont environ égaux à la moitié de ce qu'ils devraient être sans erreur de mesure. Cependant, nous trouvons que la discrétisation des mesures n'a qu'un impact très marginal sur le biais des estimateurs. Les simulations confirment également que les erreurs de mesures sont plus problématiques en ce qui concerne la significativité des coefficients lorsque les échantillons sont de petites tailles (par exemple 100 observations). La méthode instrumentale ORIV (Obviously Related Instrumental Variables) résout en moyenne le problème du biais d'atténuation mais ne résout pas complètement le problème de la significativité : les coefficients apparaissent souvent comme non significatifs alors

qu'ils devraient l'être. L'utilisation de la méthode ORIV couplée à une taille d'échantillon plus importante résout tous les problèmes.

Chapitre 4

Le chapitre 4 aborde la question du bruit dans les mesures de confiance interpersonnelle. Alors qu'une littérature toujours plus abondante évalue l'impact de la confiance sur les variables économiques et constate, par exemple, que la confiance est associée à des revenus et une croissance plus élevés, le problème potentiel de l'erreur de mesure dans les tâches déterminant la confiance des agents économiques n'a que très peu été abordé. En utilisant un protocole test-retest (une même tâche est effectuée plusieurs fois au sein d'une même expérience), nous cherchons à estimer une borne inférieure de la variance des erreurs de mesure dans deux tâches de confiance : le comportement dans un jeu de confiance et les indices de confiance déclaratifs. En outre, nous recherchons les facteurs potentiels qui pourraient jouer sur la quantité de bruit dans les tâches expérimentales. Nous soupçonnons que l'attention et la capacité cognitive peuvent jouer un rôle, comme cela a été suggéré dans des études précédentes. Nous utilisons des tâches de distraction pour obtenir une mesure de l'effort (ou de l'attention) et de la capacité cognitive.

Résultats: Nous constatons que la mesure de confiance basée sur la question d'enquête ne comporte pratiquement aucun bruit, tandis que la mesure de confiance basée sur le jeu de confiance comporte un bruit substantiel. Nous trouvons environ 15% d'erreur de mesure. Compte tenu de la spécificité de notre groupe de sujets (étudiants des meilleures écoles d'ingénieurs) et de la courte période entre les mesures test-retest, nous considérons cette valeur comme une borne inférieure. Nous constatons que les deux mesures de confiance (confiance déclarée et comportement dans le jeu de confiance) sont positivement et significativement corrélées. Enfin, en effectuant une analyse en sous-groupes, nous constatons que les sujets qui sont plus compétents sur le plan cognitif et plus attentifs prennent des décisions moins bruitées dans le jeu de confiance et dans la mesure d'aversion au risque. Néanmoins, les compétences cognitives et l'attention ne jouent aucun rôle dans le peu de bruit que comporte la mesure d'enquête.

Contents

General Introduction

This dissertation contributes to the vast area of research in behavioral and experimental economics. In this introduction, I first offer general insights into the field of research before focusing on two topics: behaviors in strategic interactions and noise in experiments. Lastly, I detail the contributions of each chapter.

While standard economics model agents as perfectly rational, some researchers – Daniel Kahneman, Amos Tversky, Richard Thaler, Vernon Smith, and Herbert Simon, to only cite a few – have challenged this assumption and have built the field of behavioral economics. Using contribution from psychology, behavioral economists have shed light on *anomalies* of human behaviors and cognitive biases that influence economic behaviors.

To understand more precisely the bounded rationality of agents, economists have used laboratory experiments. In those lab experiments, subjects perform tasks with monetary incentives. The strength of lab experiments is the environment entirely being controlled by the experimenter. As in clinical trials, experimental economists use random assignments to treatments when searching for causal inference.

Behavioral and experimental economists have exposed a list of facts that contradict standard economic theory by conducting lab experiments and analyzing field evidence. For instance, Kahneman and Tversky (1979) have exposed the asymmetric manner agents perceive financial gains and losses. Thaler (1985) laid out the mental accounting phenomenon: the fact that the way agents use their money depends on the way they earned it, whereas it should not according to standard theory. To give a last example, Kahneman et al. (1986) have shown that agents have other-regarding preferences (such as altruism, inequity aversion) while standard economics does not account for it.

One thing is to find *anomalies*, another is to build models or new theoretical settings that can account for these *anomalies* and that can compete with standard economic models. Thaler (1994) mentions this as the pursuit of a quasi-rational model (see Clochard et al., 2018, for a discussion about Thaler's vision about a Quasi-Rational model). Behavioral economists have been continuously proposing and developing behavioral models. The prospective theory (Kahneman and Tversky, 1979) and the theory of fairness, competition, and cooperation (Fehr and Schmidt, 1999) are examples of quasi-rational models that have successfully accounted for behaviors in the lab.

An entire part of behavioral and experimental economics has thus been constituted

as a mix of two things. On one side, the development of theoretical models whose origins can be introspection or empirical evidence from the lab and the field. On the other side, the design of new experiments to either validate, invalidate or calibrate theoretical models. In brief, it forms a field of science that operates back and forth between theory and experiments. I am glad to contribute to this field, developing theoretical settings and making use of laboratory experiments.

Another agenda of research in behavioral and experimental economics has consisted of measuring individual behavioral characteristics and identifying patterns and correlations between those lab behaviors and between lab behaviors and field economic decisions. Indeed, experimentalists have identified many behaviors that could correlate with economic decisions: risk-aversion, ambiguity aversion, loss aversion, inequity aversion, altruism, trust, cognitive skills, time preferences, and overconfidence, to cite the main ones. Since the nineties, experimentalists have built many tasks to elicit those behaviors. Berg et al. (1995) and Gneezy and Potters (1997) are examples of papers introducing those tasks for trust and risk preferences, respectively.

Regarding the research on the correlation between lab behaviors, we can cite three central studies. Burks et al. (2009) and Dohmen et al. (2010) both find that higher cognitive skills are associated with lower aversion to taking risks and to more patience. Chapman et al. (2018) have conducted a thorough analysis of the correlation between 21 behavior regularities on a representative sample of the American population. They find that those lab behaviors can be classified into 6 clusters of high correlations confirming previous findings. These works clarify to what extent agents are heterogeneous. Other studies use lab behaviors either as control variables or as explanatory variables to explain other outcomes in tasks in the lab that replicate economic situations on the field, such as participation in tournaments, competitions, public good games (see for instance Niederle and Vesterlund, 2007; Teyssier, 2012). They indeed find that those individual characteristics are good predictors of outcomes inside the lab.

Nonetheless, if those findings were not relevant regarding field economic decisions – in other words, if they had poor external validity– the contribution would be of lower importance. That is why, in parallel, researchers study how these lab behaviors predict decisions in the real world. For instance, the works of Eckel et al. (2005), Ashraf et al. (2006) and Burks et al. (2012) show how time preferences can predict savings or other economic outcomes in the field. Regarding risk and ambiguity preferences, Liu (2013), and Ward and Singh (2015) show that risk-aversion correlates with the adoption of new agricultural technologies in China and India, respectively. On the contrary, Charness et al. (2020) show that risk-attitude in the lab does not predict risk-attitude outside the lab using a representative sample of the Dutch population.

These approaches are not necessarily model-based. Researchers perform empirical

linear regressions or compute empirical correlations. As I will precise below, one potential issue with these approaches is measurement error that can bias coefficient in regressions and empirical correlations in general. Indeed, economic agents often exhibit noisy behaviors. The decisions made by agents, particularly during lab experiments, can be inconsistent and, for instance, influenced by the mood or recent experience. This thesis highlights and calibrates the extent of noise in experimental measures and therefore contributes methodologically to this literature.

Behavioral and Experimental Economics fields have become bigger and bigger. Nowadays, it is not surprising to introduce behavioral topics in microeconomics classes, even introductory courses. Virtually all universities have experimental laboratories in which students are the main subjects. The dissertation contributes to this large field and focuses on two main topics in behavioral and experimental economics that I will describe below more precisely: behaviors in strategic environments and noise in experimental measures.

Behaviors in strategic interactions

The first main topic addressed in this manuscript is the behaviors of agents in games. Many interactions in professional or personal life can be modeled as games, e.g., auctions, financial markets, bargaining. Cournot (1838) is the first to build a theory of strategic interactions, explicitly applied to competition between firms. More than a century later, Von Neumann and Morgenstern (1944) studied theoretically and specifically two-person zero-sum games. Nash (1951) built a unified theoretic setting modeling strategic interactions and stating how fully rational agents should behave in strategic interactions. Solutions to non-cooperative games are named after him: Nash equilibria. The concept of Nash equilibrium is still the reference when it comes to modeling strategic interactions. Nonetheless, behaviors in laboratory experiments strongly deviate from Nash theory. Indeed, when subjects are asked to play one-shot games, they often choose strategies that are not supported by Nash theory (see for instance Goeree and Holt, 2001; Camerer, 2003). While an agenda of research focuses on how other-regarding preferences may lead agents to deviate from pure rationality in games (see Cooper and Kagel, 2016, for a survey) we will focus in this dissertation on situations in which other-regarding preferences should not play a role. In this regard, two research agendas stem from the fact that subjects deviate from Nash equilibrium in lab experiments.

Promoting rationality

A strand of literature has studied mechanisms that may work as a rationality enhancer in the field, making economic agents behave closer to Nash theory. Two main mechanisms

that may be field relevant have been investigated in the lab: *learning* and *selection*. Indeed, studying learning is crucial since many economic transactions are taking place among participants who already learned through experience (List, 2003, 2004). Regarding selection, economic decision-makers may be different from the general population since they could have been selected according to different criteria (cognitive skills, experience) that may be linked to economic success. For instance, recruiters may choose candidates who obtained a specific diploma or scored well in a cognitive test. Furthermore, the market can select economic agents by eliminating participants who get bankrupt.

Researchers have explored how the *learning* mechanism can lead agents to converge to Nash equilibrium plays. Empirically, it has been shown that when agents receive feedback in repeated interactions, they converge to Nash equilibrium (see for instance Nagel, 1995; Weber, 2003; Gill and Prowse, 2016). Theoretical models were built to show why and how learning can lead to convergence to Nash equilibrium. Then, when strategic interactions are repeated, we may care less about the initial deviation to Nash from subjects in lab experiments (see Fudenberg et al., 1998; Camerer, 2003). Two limits appear. Firstly, depending on the structure of some games, learning is very long or may not emerge. Secondly, some interactions in the real world are not necessarily repeated, with or without feedback. Therefore, we may need to understand or model initial responses in games.

The *selection* mechanism has been explored by attempting to make lab subjects resemble decision-makers in the field and assessing whether they play closer to Nash equilibrium. Fréchette (2011) collected evidence from thirteen experiments involving professionals and students. Apart from one experiment, professionals and students seem to act the same way in strategic interactions and deviate from Nash equilibrium. Levitt et al. (2010) also compares the behaviors of professional poker, bridge, and soccer players with students and do not find that professionals play closer to Nash in normal form games. So far, this mechanism has not proved very efficient at filtering irrationality in the lab but may be very efficient in the field.

We contribute to this strand of literature by testing a new mechanism that may promote rationality in strategic environments: *self-selection*. Indeed, in the field, agents may self-select into situations in which they feel at ease. Therefore, behaviors in environments where self-selection is possible may be more rational and closer to the theory. Chapter 1 investigates the role of self-selection in strategic environments[1].

Behavioral game theory

Another strand of literature has built behavioral game theory and has tried to model bounded rationality in games. Two main models were elaborated in the mid-1990s namely the Quantal Response Equilibrium (QRE) model McKelvey and Palfrey (1995) and the

[1]We provide detail about the contributions of each chapter at the end of the introduction

level-k model (Stahl II and Wilson, 1994; Stahl and Wilson, 1995; Nagel, 1995).

In the QRE model, players tremble when choosing the best strategy. Players make mistakes when choosing strategies but choose low payoff strategies with low probabilities such that costly errors are not frequent. In this model, players form correct beliefs, but instead of best-responding, they *better respond*. The most common parametric form of the QRE is the Logit-QRE, in which errors follow a Gumbel distribution, and thus the difference of errors follows a centered logistic distribution. Logit-QRE model is a one-parameter model that fits empirical data well (see Goeree et al., 2016, that reviews theoretical and empirical results related to QRE model). However, one may question the micro-foundation of this common knowledge of noise in actions. Variants of the QRE models were introduced for 25 years (Weizsäcker, 2003; Goeree and Holt, 2004; Rogers et al., 2009; Goeree et al., 2018; Friedman, 2020) to better account for empirical behaviors and to offer theoretical alternatives that describe better what is in the mind of players.

In the level-k model, players are classified according to their depth of reasoning, their *level*. Level-0 players choose random strategies (or salient strategies depending on specifications) while Level-k players best respond to Level-(k-1) strategy. In the cognitive hierarchy model (Camerer et al., 2004), a variant of the level-k model, players best respond to a mixture of lower-level players. In general, to parameterize level-k models, we assume that the distribution of levels is a Poisson with parameter λ. Other variants have also been introduced (e.g., Alaoui and Penta, 2016; Koriyama and Ozkes, 2021). Level-k models have also proved to be efficient at explaining out-of-equilibrium behaviors (see, for instance, Costa-Gomes et al., 2009; Crawford et al., 2013) and capture the heterogeneity of subjects we observe in the lab.

Although many behavioral models that account for out-of-equilibrium plays have emerged since the mid-1990s, there is no clear consensus regarding the best alternative to the Nash solution concept. The fact that most theoretical papers in microeconomics still model strategic interactions with the Nash Equilibrium concept speaks for itself. Therefore, I believe that there is still room for improvement regarding the introduction of behavioral models. A model has to account for systematic empirical patterns, have interesting micro-foundations, and compete with existing models when it comes to fitting empirical strategies to be convincing. Chapter 2 contributes to this research agenda by introducing a new model, the better beliefs model.

On the typology of players

The heterogeneity of empirical behaviors has led researchers to consider that players belong to different types. For instance, in level-k models, the type of a player is its level. Many attempts were conducted to check if types of players are linked to cognitive skills and if types are stable across games. Burnham et al. (2009) Carpenter et al. (2013)

finds that higher cognitive skills measured by IQ or Cognitive Reflection Test (CRT, see Frederick, 2005) score are associated with higher strategic sophistication (higher levels of reasoning) in strategic games. Brañas-Garza et al. (2012) finds that Raven's Matrices test (Raven, 1936) score does not correlate with higher strategic sophistication while CRT does. Fehr and Huck (2016) finds the same result regarding CRT and strategic sophistication. Georganas et al. (2015) finds that IQ, CRT, and Raven Test scores only poorly correlate with levels of reasoning. Gill and Prowse (2016) finds that subjects with higher CRT scores are not significantly more strategic in the first repetition of a beauty contest game. Nonetheless, they seem to learn faster. Overall, results are ambiguous regarding the link between cognitive skills and types measured as strategic sophistication (e.g., level-k), and no conclusion can be drawn.

Regarding the stability of types, Georganas et al. (2015) find that, in games where level-k types are identifiable, they are not stable at all across families of games. Cooper et al. (2016) also finds that individual behaviors cannot be explained with a fixed level of reasoning. Rubinstein (2016) built a typology of players with two types: instinctive and contemplative. These types correlate with response times of players; the longer the response times, the more contemplative the strategy on average. A Contemplative Index built with individual behaviors in 10 games correlates with contemplative answers with other games but only very poorly. Results regarding the stability of types seem to show that as psychologists Gick and Holyoak (1980), Perkins and Salomon (1989) explained, expertise can be context-specific. Agents can be successful, strategically sophisticated in one game but not necessarily in another.

Chapter 1 fuels the ongoing debate on the stability of types of players by introducing a measure of confidence which, to the best of our knowledge, is new in the context of strategic interactions. The model in Chapter 2 introduces types of players but does not necessarily impose the stability of those types across games.

Noise in experiments

The other topic tackled in the dissertation is noise in experimental measures. Measurement uncertainty has always been an issue for scientists (see Stigler, 1986, for a thorough account of the measurement of uncertainty before the nineteenth). A striking example of scientists dealing with measurement error is the difficulties encountered by the brilliant mathematician Leonhard Euler. Considering imprecise observations as actual parameters, Euler (1749) struggled to explain theoretically the movement of Saturn and Jupiter planets: "Now from these equations we can conclude nothing, perhaps, is that I have tried to satisfy several observations exactly, whereas I should have only satisfied them approximately; and this error has then multiplied itself.". On the contrary, at the same

period, Tobias Mayer (Mayer, 1750) successfully used observational data, averaging linear equations to combine empirical observations to explain the Moon's libration. What was only heuristics in the 18^{th} century became proved theoretically in the 19^{th} century with the generalization of Ordinary Least Squares (OLS) methods. Nonetheless, OLS regressions do not solve all statistical issues, in particular when dependant variables are measured with error. Indeed, we know at least from the work of Spearman (1904) that correlations and coefficients in OLS regressions can be biased if variables are measured with error. Spearman is the first to refer to *attenuation bias* the fact that the correlation between variables measured with error is lower in absolute value. We provide below a simple insight into how OLS estimates are biased.

Consider the following model in which an outcome y has a linear relation with a variable x^*:

$$y = \alpha + \beta x^* + \eta \quad \text{with} \quad cov(x^*, \eta) = 0^2$$

Consider that we only observe a noisy measure x of x^*:

$$x = x^* + \epsilon \quad \text{with} \quad cov(x^*, \epsilon) = 0 \quad \text{and} \quad cov(\epsilon, \eta) = 0$$

If we regress the outcome y on the noisy measure x the OLS estimator $\hat{\beta}$ satisfies:

$$\hat{\beta} \xrightarrow{n \to +\infty} \frac{cov(y, x)}{var(x)} = \frac{var(x^*)}{var(x^*) + var(\epsilon)} \beta$$

We observe in this simple example that $\hat{\beta}$ is biased and that the estimated effect is an attenuation (in absolute value) of the true effect.

More than a century later, Gillen et al. (2019) deal with the specific problem of measurement error in laboratory experiments and proposes an instrumental approach denoted as ORIV (Obviously Related Instrumental Variables). They argue that errors in measurement are ubiquitous in experiments while rarely taken into account. As a matter of fact, when we ask subjects to perform twice the same task in a short period of time, we often obtain different choices. Econometricians model this within-individual variability as noise or measurement error. The source of measurement error in experiments can be diverse, e.g., inattention, imprecise preferences, or mood's influence. Recently, Enke and Graeber (2019) have introduced the concept of *cognitive uncertainty* to refer to the noise affecting the decision process. Calibrating measurement error in experimental tasks is very important to assess the magnitude of the attenuation bias regressions or correlations could suffer from. Furthermore, when estimating multivariate OLS regressions, the coefficient of variables measured without error can be biased if one variable included in the regression is

[2] *cov* represents the theoretical covariance. We furthermore consider that all necessary conditions of moments regarding the convergence of OLS estimators are satisfied.

measured with error. Gillen et al. (2019) show, for instance, that in some studies, a gender variable appears significant while it should not if preferences such as risk-attitude were measured without errors (or if instrumental methods were used). Errors in measurement are thus very problematic, and we address this issue in the dissertation by focusing on two behavioral regularities, risk-attitude and trust. Two chapters are devoted to calibrating and discussing errors in measurement in those tasks. These methodological discussions are of prime importance since, as highlighted above, a whole field of experimental economics uses behavioral regularities – that can be measured with error – to explain or predict outcomes in or outside the lab. Therefore, those two chapters can help raise awareness about the issue of measurement error.

Chapter 1

Chapter 1 examines a mechanism that could promote rationality in strategic environments: self-selection. It starts from a simple observation. In most experiments, once seated in the lab, subjects are forced to participate in all the tasks they faced. On the contrary, in many economic situations (e.g., financial markets and auctions), agents can choose whether they want to participate. This self-selection stage that occurs in the field may influence the behaviors in economic situations. We design an experiment with an explicit self-selection stage and assess the evolution of strategies. We furthermore study the potential drivers of self-selection into games. We complete our analysis by reviewing existing evidence in the field of self-selection into strategic environments. Finally, we discuss how self-selection could improve the external validity of lab experiments.

Results: We find that strategies do significantly evolve under self-selection. Self-selection act as a filter of irrationality. For instance, fewer dominated strategies are played, the proportion of players choosing Nash strategies increases, and subjects' strategic sophistication is higher. However, players do not adapt much to changes in the subject pool resulting from self-selection. Regarding the drivers of self-selection, we find that risk-averse players tend to enter less in strategic environments. In addition, a confidence measure we introduced, which is new in the context of experimental games, correlates with the entry decision. Furthermore, this confidence measure is game-specific and well captures a sense of doing the right thing in a game since it correlates well with the strategies played in the game in question. Our review of field evidence leads to similar patterns: self-selected economic agents are less risk-averse, more rational, and more homogeneous. We hope that this study will generate interest regarding the self-selection mechanism that could be implemented in many experimental settings.

This chapter circulates as a working paper (Hollard and Perez, 2020).

Chapter 2

Chapter 2 introduces a new model in behavioral game theory. The model accounts for two robust empirical facts: the existence of non-strategic subjects and the heterogeneity in beliefs. Indeed, half a dozen papers (Costa-Gomes and Weizsäcker, 2008; Agranov et al., 2012; Burchardi and Penczynski, 2014; De Sousa et al., 2013; Agranov et al., 2015; Fragiadakis et al., 2016) have identified, with different methodologies, a significant proportion (between 30% and 50%) of confused subjects that are not able to form beliefs. Regarding heterogeneity in beliefs, Costa-Gomes and Weizsäcker (2008) and Fragiadakis et al. (2019), for instance, show that stated beliefs greatly vary across subjects. There are two types of players in the model we introduce: not-forming beliefs players and strategic players. As their denomination suggests, not-forming beliefs (NFB) players cannot form any beliefs and play randomly or naively. On the contrary, strategic beliefs do form beliefs on the probability of facing a not-forming beliefs player or a strategic player as sophisticated as herself. Nonetheless, strategic players have noisy (but still close to perfect) beliefs regarding the proportion of NFB and strategic players. Strategic players best respond to their heterogeneous beliefs but ignore that other strategic players may have different beliefs.

Results: We introduce the better beliefs model. We show the existence of a one-parameter version of the model in which strategic agents have heterogeneous beliefs regarding the opponents they face, thanks to semi-algebraic geometry arguments. This model has interesting theoretical properties. In particular, when the parameter goes from one limit to another, the distribution of predicted strategies goes from random to Nash Equilibrium strategies. The parameter directly depicts how strategic a subject's pool is. Furthermore, the model fits and correctly predicts experimental data from two previous studies.

This chapter circulates as a working paper (Hollard and Perez, 2022).

Chapter 3

Chapter 3 assesses the measurement error problem in typical experimental tasks, namely risk-aversion tasks. Eliciting risk-aversion is important since many economic decisions in life may be linked to individual relative risk aversion: real estate investment, health insurance, life insurance, job positions. Therefore, assessing how risk-averse agents are may be important to predict or explain economic behaviors. Risk-aversion tasks were thus designed (see for instance Holt and Laury, 2002; Crosetto and Filippin, 2013). In this chapter, we try to calibrate the part of noise entailed in those tasks since, as mentioned above in the introduction, measurement error can have dramatic statistical consequences.

We first calibrate the part of errors in measurement using five different studies with four different tasks and 16 different datasets. We perform simulations to assess the impact of measurement error coupled with a small sample on the significance of coefficients. We also try to determine whether the fact that most measures of risk-aversion are discrete and not continuous can be problematic when looking for correlations between variables.

Results: We find that risk-aversion measures entail a significant measurement error component: between 35% and 60% of the variance of the measure is due to measurement error. As the theory predicts, this leads to an attenuation bias of about one-half: empirical estimates of coefficients in our simulation of simple OLS regressions approximately equals half of what it should be without measurement error. However, we find that the discretization of the measures has only a marginal effect on biasing coefficients. Simulations also confirm that measurement error is more problematic regarding significance when coupled with a small sample. Indeed, empirical coefficients are often non-significant when we have, for instance, only 100 observations. ORIV method is found to solve the attenuation bias issue on average but does not fully solve the significance issue: many estimates still appear non-significant when they should. Simulations confirm that using ORIV estimations coupled with a larger sample size solves all the issues.

This chapter is published in the Journal of Risk and Uncertainty (Perez et al., 2021).

Chapter 4

Chapter 4 addresses the issue of noise in measures of trust. While ever-growing literature assesses the impact of trust on economic variables and finds, for instance, that trust is associated with higher income and growth, the potential issue of measurement error in tasks eliciting trust has never been discussed. Only Chapman et al. (2018) used the ORIV method to instrument the behavior of the receiver in a trust game to correct for measurement error. Using a test-retest protocol, we are looking to estimate a lower bound of the variance of the errors in measurement in two trust tasks: stated trust (non-incentivized survey question) and sender behavior in a trust game. Furthermore, we search for potential drivers of noise in experiments. We suspect that attention and cognitive ability may play a role as it has been suggested by Anderson and Mellor (2009) and Amador-Hidalgo et al. (2021). We use the distractive tasks to obtain a measure of effort (or attention) and cognitive ability.

Another objective is to estimate the correlation between stated trust and trust behavior in a trust game and show how correcting for measurement error can improve correlation estimation. Whether those two tasks measure a common behavior (or

characteristic) is still an ongoing debate in the literature, and previous studies (Glaeser et al., 2000; Lazzarini et al., 2003; Fehr et al., 2003; Ashraf et al., 2004; Bellemare and Kröger, 2007; Falk et al., 2016; Thöni et al., 2012; Sapienza et al., 2013; Banerjee, 2018) have shown ambiguous results.
(Glaeser et al., 2000; Lazzarini et al., 2003; Fehr et al., 2003; Ashraf et al., 2004; Bellemare and Kröger, 2007; Falk et al., 2016; Thöni et al., 2012; Sapienza et al., 2013; Banerjee, 2018) have shown ambiguous results.

Results: We find that the trust measure based on the survey question entails virtually no noise, while the trust measure based on the trust game does entail substantial noise. We find about 15% of measurement error. Given the specificity of our subject pool (students in top engineering schools) and the short time between test-retest measures, we consider this value as a lower bound. We find that both trust measures (stated trust and behavior in the trust game) are positively and significantly correlated and also correlate with trustworthiness. Since measurement error is relatively low, correcting for measurement error does not change much the value of empirical correlations. Lastly, by performing a subgroup analysis, we find that subjects who are more cognitively skilled and more attentive exhibit less noisy decisions in the trust game and the risk-aversion measure. Nonetheless, cognitive skills and attention do not play any role in the little noise entailed in the survey measure. We suggest that there could exist a trade-off between incentive and less noise in experimental tasks.

Note

The four chapters of this dissertation are independent research articles. This is why some information may be redundant and why the term *article* is sometimes used instead of *chapter*. Chapter 1 and Chapter 2 are co-authored with Guillaume Hollard. I am first author of Chapter 3, with Guillaume Hollard and Radu Vranceanu being co-authors. Chapter 4 is co-authored with Héloïse Clolery, Guillaume Hollard and Inès Picard.

References

Agranov, M., A. Caplin, and C. Tergiman (2015). Naive play and the process of choice in guessing games. *Journal of the Economic Science Association 1*(2), 146–157.

Agranov, M., E. Potamites, A. Schotter, and C. Tergiman (2012). Beliefs and endogenous cognitive levels: An experimental study. *Games and Economic Behavior 75*(2), 449–463.

Alaoui, L. and A. Penta (2016). Endogenous depth of reasoning. *The Review of Economic Studies 83*(4), 1297–1333.

Amador-Hidalgo, L., P. Brañas-Garza, A. M. Espín, T. García-Muñoz, and A. Hernández-Román (2021). Cognitive abilities and risk-taking: Errors, not preferences. *European Economic Review 134*, 103694.

Anderson, L. and J. Mellor (2009, 09). Are risk preferences stable? comparing an experimental measure with a validated survey-based measure. *Journal of Risk and Uncertainty 39*, 137—160.

Ashraf, N., I. Bohnet, and N. Piankov (2004). Is trust a bad investment? *SSRN Electronic Journal*.

Ashraf, N., D. Karlan, and W. Yin (2006). Tying odysseus to the mast: Evidence from a commitment savings product in the philippines. *The Quarterly Journal of Economics 121*(2), 635–672.

Banerjee, R. (2018). On the interpretation of world values survey trust question-global expectations vs. local beliefs. *European Journal of Political Economy 55*, 491–510.

Bellemare, C. and S. Kröger (2007). On representative social capital. *European Economic Review 51*(1), 183–202.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity, and social history. *Games and economic behavior 10*(1), 122–142.

Brañas-Garza, P., T. García-Muñoz, and R. H. González (2012). Cognitive effort in the beauty contest game. *Journal of Economic Behavior & Organization 83*(2), 254–260.

Burchardi, K. B. and S. P. Penczynski (2014). Out of your mind: Eliciting individual reasoning in one shot games. *Games and Economic Behavior 84*, 39–57.

Burks, S., J. Carpenter, L. Götte, and A. Rustichini (2012). Which measures of time preference best predict outcomes: Evidence from a large-scale field experiment. *Journal of Economic Behavior & Organization 84*(1), 308–320.

Burks, S. V., J. P. Carpenter, L. Goette, and A. Rustichini (2009). Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences 106*(19), 7745–7750.

Burnham, T. C., D. Cesarini, M. Johannesson, P. Lichtenstein, and B. Wallace (2009). Higher cognitive ability is associated with lower entries in a p-beauty contest. *Journal of Economic Behavior & Organization 72*(1), 171–175.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics 119*(3), 861–898.

Carpenter, J., M. Graham, and J. Wolf (2013). Cognitive ability and strategic sophistication. *Games and Economic Behavior 80*, 115–130.

Chapman, J., M. Dean, P. Ortoleva, E. Snowberg, and C. Camerer (2018). Econographics. Technical report, National Bureau of Economic Research.

Charness, G., T. Garcia, T. Offerman, and M. C. Villeval (2020). Do measures of risk attitude in the laboratory predict behavior under risk in and outside of the laboratory? *Journal of Risk and Uncertainty 60*(2), 99–123.

Clochard, G.-J., G. Hollard, and F. Perez (2018). Richard h. thaler et les limites de la rationalité. *Revue d'economie politique 128*(4), 535–548.

Cooper, D., E. Fatas, A. J. Morales, and S. Qi (2016). Consistent depth of reasoning in level-k models. Technical report, Technical report.

Cooper, D. J. and J. H. Kagel (2016). 4. other-regarding preferences. In *The Handbook of Experimental Economics, Volume 2*, pp. 217–289. Princeton University Press.

Costa-Gomes, M. A., V. P. Crawford, and N. Iriberri (2009). Comparing models of strategic thinking in van huyck, battalio, and beil's coordination games. *Journal of the European Economic Association 7*(2-3), 365–376.

Costa-Gomes, M. A. and G. Weizsäcker (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies 75*(3), 729–762.

Cournot, A. A. (1838). *Recherches sur les principes mathématiques de la théorie des richesses*. L. Hachette.

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature 51*(1), 5–62.

Crosetto, P. and A. Filippin (2013). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty 47*(1), 31–65.

De Sousa, J., G. Hollard, and A. Terracol (2013). Non-strategic players are the rule rather than the exception. Technical report, working paper, Université Paris 1.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review 100*(3), 1238–60.

Eckel, C., C. Johnson, and C. Montmarquette (2005). *Saving decisions of the working poor: Short-and long-term horizons*. Emerald Group Publishing Limited.

Enke, B. and T. Graeber (2019). Cognitive uncertainty. Technical report, National Bureau of Economic Research.

Euler, L. (1749). Recherches sur la question des inégalités du mouvement de saturne et de jupiter. *Pièce qui ont remporté le prix de l'académie royale des sciences*, 1–123.

Falk, A., A. Becker, T. J. Dohmen, D. Huffman, and U. Sunde (2016). The preference survey module: A validated instrument for measuring risk, time, and social preferences.

Fehr, D. and S. Huck (2016). Who knows it is a game? on strategic awareness and cognitive ability. *Experimental Economics 19*(4), 713–726.

Fehr, E., U. Fischbacher, J. Schupp, B. Rosenbladt, and G. Wagner (2003, 05). A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative surveys. *Schmoller's Jahrbuch 122*.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics 114*(3), 817–868.

Fragiadakis, D. E., D. T. Knoepfle, and M. Niederle (2016). Who is strategic? Technical report, Working Paper. 1.1.

Fragiadakis, D. E., A. Kovaliukaite, and D. R. Arjona (2019). Belief-formation in games of initial play: an experimental investigation.

Fréchette, G. R. (2011). Laboratory experiments: Professionals versus students. *Available at SSRN 1939219*.

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

Friedman, E. (2020). Endogenous quantal response equilibrium. *Games and Economic Behavior 124*, 620–643.

Fudenberg, D., F. Drew, D. K. Levine, and D. K. Levine (1998). *The theory of learning in games*, Volume 2. MIT press.

Georganas, S., P. J. Healy, and R. A. Weber (2015). On the persistence of strategic sophistication. *Journal of Economic Theory 159*, 369–400.

Gick, M. L. and K. J. Holyoak (1980). Analogical problem solving. *Cognitive psychology 12*(3), 306–355.

Gill, D. and V. Prowse (2016). Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis. *Journal of Political Economy 124*(6), 1619–1676.

Gillen, B., E. Snowberg, and L. Yariv (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy 127*(4), 1826–1863.

Glaeser, E. L., D. I. Laibson, J. A. Scheinkman, and C. L. Soutter (2000). Measuring trust. *The quarterly journal of economics 115*(3), 811–846.

Gneezy, U. and J. Potters (1997). An experiment on risk taking and evaluation periods. *The quarterly journal of economics 112*(2), 631–645.

Goeree, J. K. and C. A. Holt (2001). Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review 91*(5), 1402–1422.

Goeree, J. K. and C. A. Holt (2004). A model of noisy introspection. *Games and Economic Behavior 46*(2), 365–382.

Goeree, J. K., C. A. Holt, and T. R. Palfrey (2016). *Quantal Response Equilibrium: A Stochastic Theory of Games*. Princeton University Press.

Goeree, J. K., P. Louis, and J. Zhang (2018). Noisy introspection in the 11–20 game. *The Economic Journal 128*(611), 1509–1530.

Hollard, G. and F. Perez (2020). Self-selection filters irrationality in one-shot games. Technical report, Center for Research in Economics and Statistics.

Hollard, G. and F. Perez (2022). Better beliefs in normal-form games. *Available at SSRN 4052270*.

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American economic review 92*(5), 1644–1655.

Kahneman, D., J. L. Knetsch, and R. H. Thaler (1986). Fairness and the assumptions of economics. *Journal of business*, S285–S300.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–292.

Koriyama, Y. and A. I. Ozkes (2021). Inclusive cognitive hierarchy. *Journal of Economic Behavior & Organization 186*, 458–480.

Lazzarini, S., R. Madalozzo, R. Artes, and J. Siqueira (2003, 11). Measuring trust: An experiment in brazil. *IBMEC Working Paper No. 01-2004*.

Levitt, S. D., J. A. List, and D. H. Reiley (2010). What happens in the field stays in the field: Exploring whether professionals play minimax in laboratory experiments. *Econometrica 78*(4), 1413–1434.

List, J. A. (2003). Does market experience eliminate market anomalies? *The Quarterly Journal of Economics 118*(1), 41–71.

List, J. A. (2004). Neoclassical theory versus prospect theory: Evidence from the marketplace. *Econometrica 72*(2), 615–625.

Liu, E. M. (2013). Time to change what to sow: Risk preferences and technology adoption decisions of cotton farmers in china. *Review of Economics and Statistics 95*(4), 1386–1403.

Mayer, T. (1750). Abhandlung über die umwälzung des monds um seine axe, und die scheinbare bewegung der mondsflecken. *Kosmographische Nachrichten und Sammlungen auf das Jahr 1748 (Part II: Kosmographische Sammlungen auf das Jahr 1748) 41*, 51.

McKelvey, R. D. and T. R. Palfrey (1995). Quantal response equilibria for normal form games. *Games and economic behavior 10*(1), 6–38.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review 85*(5), 1313–1326.

Nash, J. (1951). Non-cooperative games. *Annals of mathematics*, 286–295.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The quarterly journal of economics 122*(3), 1067–1101.

Perez, F., G. Hollard, and R. Vranceanu (2021). How serious is the measurement-error problem in risk-aversion tasks? *Journal of Risk and Uncertainty*, 319–342.

Perkins, D. N. and G. Salomon (1989). Are cognitive skills context-bound? *Educational researcher 18*(1), 16–25.

Raven, J. C. (1936). Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive. *Unpublished master's thesis, University of London.*

Rogers, B. W., T. R. Palfrey, and C. F. Camerer (2009). Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory 144*(4), 1440–1467.

Rubinstein, A. (2016). A typology of players: Between instinctive and contemplative. *The Quarterly Journal of Economics 131*(2), 859–890.

Sapienza, P., A. Toldra-Simats, and L. Zingales (2013). Understanding trust. *The Economic Journal 123*(573), 1313–1332.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology 15*(1), 72–101.

Stahl, D. O. and P. W. Wilson (1995). On players models of other players: Theory and experimental evidence. *Games and Economic Behavior 10*(1), 218–254.

Stahl II, D. O. and P. W. Wilson (1994). Experimental evidence on players' models of other players. *Journal of economic behavior & organization 25*(3), 309–327.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Harvard University Press.

Teyssier, S. (2012). Inequity and risk aversion in sequential public good games. *Public Choice 151*(1), 91–119.

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing science 4*(3), 199–214.

Thaler, R. H. (1994). *Quasi rational economics.* Russell Sage Foundation.

Thöni, C., J.-R. Tyran, and E. Wengström (2012). Microfoundations of social capital. *Journal of Public Economics 96*(7-8), 635–643.

Von Neumann, J. and O. Morgenstern (1944). Theory of games and economic behavior.

Ward, P. S. and V. Singh (2015). Using field experiments to elicit risk and ambiguity preferences: Behavioural factors and the adoption of new agricultural technologies in rural india. *The Journal of Development Studies 51*(6), 707–724.

Weber, R. A. (2003). 'learning'with no feedback in a competitive guessing game. *Games and Economic Behavior 44*(1), 134–144.

Weizsäcker, G. (2003). Ignoring the rationality of others: evidence from experimental normal-form games. *Games and Economic Behavior 44*(1), 145–171.

Chapter 1

Self-Selection as a Filter of Irrationality in Games

Note: This chapter is co-authored with Guillaume Hollard.[1].

Abstract

If economic agents engaging in strategic interactions have a sense of how well they are doing, they should avoid situations in which they may make embarrassing mistakes by opting out. We here propose to test that proposition and to explore its consequences regarding self-selection into strategic interactions. A crucial possible effect is that strategic interactions among self-selected agents (which are common in the field) may be better described by standard theory than situations in which agents did not have the opportunity to self-select (like standard experiments in the lab). We here investigate the effect of self-selection on the outcome of games. We propose an experiment in which subjects can (1) report their subjective feeling of how well they are doing using a confidence measure that is new in this context and (2) self-select into strategic interactions or choose a sure payoff. We find that self-selection is indeed driven by our measure of confidence and risk-aversion: the least confident and more risk-averse subjects are less likely to self-select into strategic interactions. Self-selection dramatically reduces the gap between theoretical predictions and actual behavior. We find that economic agents do have a sense of how well they are doing in games. Perhaps more surprisingly, this sense varies significantly from game to game at the individual level. We complement our analysis by reviewing field evidence on self-selection, which confirms that self-selection promotes rationality.

Keywords: Self-selection; Experiments; Non-cooperative games; Rationality.

[1]For this chapter I specifically would like to thank John List, Marie-Claire Villeval and Georg Weizsäcker for insightful comments and encouragements. I also thank Marina Agranov, Wael Bousselmi, Lawrence Choo, Gwen-Jiro Clochard, Philippe Choné, Jules Depersin, Nobuki Hanaki, Yukio Koryama, Dorothea Kubler, Al Roth, Angela Sutan, Anaëlle Touré, Pedro Vergara Merino, as well as participants at the 2019 ESA Europe Meeting, the 2019 ASFEE conference, the 2019 ESA AP Meeting and the 2019 CWEE workshop at CREST

1. Introduction

Let us consider an economic agent who is contemplating the possibility of engaging in a strategic interaction. If she feels she may make embarrassing mistakes, she may opt out. As a result, when players are offered the possibility to self-select, we expect the least rational ones to opt out. In other words, we here suggest that economic agents can anticipate their relative ability to act strategically. The mere fact that many economic interactions take place among *self-selected* subjects may thus be a powerful driver of rationality. In particular, strategic interactions among self-selected agents (which are common in the field) may be better described by standard theory than situations in which agents did not have the opportunity to self-select (like in standard experiments in the lab).

While many aspects of self-selection have been studied,[1] few studies are comparing the difference in rationality (or strategic sophistication) between self-selected and non-self-selected subject pools. Also, there is still little evidence on whether agents are able to anticipate their own ability to play games.

So far, two main forces have been proved to promote rationality in strategic environments: learning (e.g., becoming expert by trial and mistakes) and selection out of the market because of too many costly mistakes. Both forces converge so that many economic interactions take place among experts. We here show that a third, hidden, force may be at work to promote rationality, namely self-selection.

To single out the effect of self-selection, we here focus on situations in which the effects of learning and selection resulting from costly mistakes have not happened yet (e.g., agents did not have a chance to get feedback), i.e., one-shot games. We use data from an original lab experiment involving one-shot games to gauge the effect of self-selection on rationality. We compare the behaviors of individuals in situations in which they had no choice but to participate to situations in which they had the choice to opt out, based on various rationality criteria. We also introduce a (new in this context) measure of confidence to elicit how subjects perceive their own ability to play games in a meaningful way. Subjective assessments are then compared to objective outcomes. Last, we review field evidence in which it is possible to compare self-selected groups of participants to non-self-selected ones.

We find that subjects are aware of their own ability to play games. Confused subjects tend to avoid opting into situations where they may make embarrassing mistakes, while those who feel at ease are more likely to self-select. We also find, as expected, that risk-averse subjects tend to avoid strategic interactions in favor of a sure payoff. Furthermore, both drivers of self-selection, risk-aversion, and confidence, are found to be uncorrelated. As a result, self-selected players better satisfy two key assumptions on which Nash equi-

[1] The review section will provide more precise definitions of what we consider as self-selection

librium is resting: risk neutrality and common knowledge of rationality. In line with this result, the outcome of the games played by self-selected subjects is found to be closer to Nash predictions. Self-selection thus increases the predictive power of Nash equilibrium. Our review of field evidence confirms these findings: self-selected agents are more rational and closer to risk neutrality. Overall, self-selection seems efficient at filtering out the most confused and more risk-averse agents.

We finally discuss in a dedicated section how self-selection could enhance external validity. While a significant part of our knowledge about how agents actually play games comes from lab experiments, we note that few experiments involving one-shot games offer subjects the opportunity to opt out of strategic interactions. In sharp contrast, self-selection is ubiquitous in the field, as individuals can choose to engage in strategic interactions or not. We hope to generate interest and open a discussion on the value of adding self-selection stages in lab experiments.

The remainder of the paper is organized as follows. Section 2 presents the literature review. Section 3 introduces the protocol of the lab experiment, and Section 4 analyses the results. Section 5 reviews evidence from the field on self-selection, and Section 6 discusses the link between self-selection and external validity. Section 7 concludes.

2. Review

We first clarify our use of the term *self-selection* and then explore existing evidence on self-selection in experimental settings. A key aspect of the present research is to link self-selection to strategic attitude in games. To this aim, we here borrow a notion of confidence from cognitive psychology that we describe below. Since we expect self-selection to promote rationality, we also briefly review alternative mechanisms impacting rationality.

2.1. Self-selection and rationality

Before reviewing existing results directly related to our question of interest, it is certainly worth clarifying how the present research takes place within existing research on self-selection.

Self-selection is often studied as a mechanism by which the considered group of individuals has been formed (e.g. individuals who self-select into specific occupations, a specific type of school, into an experimental treatment, etc). A key information is whether, and to what extent, considered groups are different from random draws from the population of interest (see Heckman, 1990).

We here focus on the particular case of self-selection that could occur once seated in the experimental room, and that would be based on the ability to choose sound strategies

in (one-shot) games. Therefore, we do not directly address other types of selection (see subsection 2.4 for more detail). We expect that players who self-select into games are different from those who do not. So, if strategic ability is heterogeneous across individuals, we expect the most strategic individuals to opt-in and the least strategic ones to opt out.

2.2. Review of experimental evidence on Self-selection

To clarify how the present study is related to others, which may use the term *self-selection* differently, we present existing evidence in a synthetic table (see Table 1.1).

Experiments on self-selection mainly study the determinants of self-selection into payment schemes. Typically, subjects can choose between a risk-free payment scheme and one involving competition, namely a tournament. Formally, a tournament involves strategic uncertainty since payments do depend on the action of others. Tournaments can thus be considered as non-cooperative games. However, the critical issue in a tournament is to anticipate one's rank within a particular subject pool. Tournaments represent a particular type of game since, in most studied situations, forming beliefs about other players consists in estimating one's rank. Players who anticipate a high ranking are likely to self-select; those who expect a low ranking are less likely to do so. In contrast, self-selection into games in general, not just tournaments, has been seldom studied.

We review all experiments we are aware of in which participants are offered the possibility to self-select into strategic interactions. Existing evidence on self-selection boils down to choosing between two payment schemes, A and B. For instance, individuals must choose between (A) being paid according to a piece rate or (B) entering a tournament. Evidence can be presented synthetically in Table 1.1. Shaded cells indicate situations in which strategic ability plays a significant role in one of the tasks at stake. A shaded cell in the last column indicates that the experimental design elicits some rationality criterion.

As can be seen, few studies are linking self-selection and strategic sophistication. There is literature on self-selection into situations in which pro-sociability plays a role, like social dilemmas. For instance, there are studies on self-selection into dictator games (Dana et al., 2006) or in the public good game (Brekke et al., 2011). We concentrate on self-selection along the strategic dimension by focusing on situations in which prosociality is unlikely to play an important role. So, we end up with three studies that are the closest to our purpose. We describe them in more detail below.

Table 1.1: Review on Self-Selection in lab experiments

		Payment Scheme A		Payment Scheme B		
Reference	Sorting	Task	Strt. Unct.	Task	Strt. Unct.	Interpretation
NV (2007) Step 3	Bin. C.	Tournament	No	Piece-rate	No	Taste for competition
NV (2007) Step 4	Bin. C.	Tournament	No	Piece-rate	No	Taste for competition
CL (1999) Treat.1	Bin. C.	Lottery	No	Sure Payoff	No	Collective ability to coordinate
CL (1999) Treat.2	Bin. C.	Tournament	No	Sure Payoff	No	Confidence in relative performance
DF (2011) Treat.A	Bin. C.	Piece-rate	No	Fixed payment	No	Self-sorting into payment schemes
DF (2011) Treat.B	Bin. C.	Tournament	No	Fixed payment	No	Self-sorting into payment schemes
DF (2011) Treat.C	Bin. C.	Revenue Sharing	No	Fixed payment	No	Self-sorting into payment schemes
DCD (2006)	Bin. C.	Dictator Game	No	Sure Payoff	No	Psychological payoffs
LMW (2012)	Bin. C.	Dictator Game	No	Sure Payoff	No	Sorting and pro-sociality
SS (2013)	Auct.	Ultim. Game	Yes	Sure Payoff	No	Sorting and pro-sociality
CIQ (2012)	Auct.	Weak link Game	Yes	Weak link Game	Yes	Coordination and efficiency
ETV (2009)	Bin. C.	Tournament	Yes	Piece-rate	Yes	Self-sorting into payment scheme
BHLN (2011)	Bin. C.	Pub. Gd Game	Yes	Pub. Gd Game	Yes	Sorting and pro-sociality
CTZ (2019)	Auct.	Beauty Cst	Yes	Sure Payoff	Yes	Sorting on strategic sophistication
CZ (2020)	Auct.	Red hat pb	Yes	Sure Payoff	Yes	Sorting on strategic sophistication
ENV (2020)	Bin. C.	Bargaining	Yes	Sure Payoff	No	Performance: Forced vs self-selected
This Paper	Bin. C.	Various Games	Yes	Sure Payoff	No	Sorting on strategic sophistication

Each line represents an experimental treatment. Shaded cells indicate that (1) one task involves strategic uncertainty and (2) the purpose of the experiment is related to strategic sophistication. Initials refers to papers: NV for Niederle and Vesterlund (2007); CL for Camerer and Lovallo (1999); DF for Dohmen and Falk (2011); DCD for Dana et al. (2006); LMW for Lazear et al. (2012); SS for Shachat and Swarthout (2013); CIQ for Cooper et al. (2018); ETV for Eriksson et al. (2009); BHLN for Brekke et al. (2011); CTZ for Choo et al. (2019); CZ for Choo and Zhou (2019); ENV for Exley et al. (2020). Column 2: Sorting is made via either a binary choice (Bin. C.) or an Auction.

Exley et al. (2020) propose an experiment that includes an explicit comparison between self-selected and not self-selected subject pools. In one treatment, subjects are forced to participate, while in the other, they are offered the choice between a sure payoff and entering a bargaining game. They find that subjects self-select based on their anticipated ability to make a profit in the subsequent bargaining game. When forced to participate, subjects end up more often with a negative payoff. So those who prefer not to self-select into games are indeed making a sensible choice on average. Furthermore, there is a difference in risk-aversion, with self-selected subjects being closer to risk neutrality.

Closer to our subject here, Choo et al. (2019) designed an auction where players bid to have the right to re-play a beauty-contest game (after having played this game against the whole group), with the eight highest bids subsequently playing the game a second time. The willingness to pay to re-play the game depends on the strategic sophistication of the players via a bonus that depends on the average chosen strategy in the beauty contest. The bonus can be high when the average is low or, on the contrary, when the average is high, making it more or less profitable for high levels to enter (where levels are defined as in the level-k model and are thus synonymous with degrees of reasoning). Subjects are found to make sensible entry decisions (i.e., they bid according to their elicited strategic sophistication, as defined in level-k models). However, the strategy they use in the subsequent beauty contest remains, on average, the same as that used in the same game without self-selection.

In a companion paper, Choo and Zhou (2019) auctioned the right to play a three-person *red hat puzzle game* and compared the issue against the control treatment where participation rights are randomly allocated. They find that auctioning the rights to participate reduces anomalous behavior by a substantial amount.

All three studies confirm that players are able to self-select based on their ability to perform in strategic interactions. Performance is indeed lower when there is no self-selection.

Furthermore, we can draw some general conclusions regarding self-selection since some consistent results across studies listed in Table 1.1 do emerge.

1. Risk-averse individuals tend to avoid strategic interactions (e.g. competitive payment schemes) and to favor riskless payments. As a result, self-selected subject pools are closer to risk neutrality.

2. Self-selection creates more homogeneous subject pools regarding criteria like risk-aversion, the type of strategy used and payoffs (see for instance Eriksson et al., 2009)

3. Agents insufficiently adjust to changes in the pool of competitor. For instance, in

Camerer and Lovallo (1999) agents fail to anticipate the consequences of facing a pool of competitors who were selected based on their anticipated performance.

4. When some kind of strategic sophistication is measured, results consistently indicate that self-selected subjects are more strategic.

The listed properties above offer insights regarding the type of results that are expected regarding the effect of self-selection.

2.3. Metacognition, Confidence, and Rationality

Metacognition is knowledge about knowledge. More precisely and closer to our purpose here, metacognition can be defined as the ability of an individual to judge the effectiveness of a particular cognitive strategy she used. Cognitive sciences have a long tradition of eliciting metacognition using *confidence judgments* (see for instance Fleming and Lau, 2014; Ais et al., 2016). In a typical experiment, subjects perform a task and are subsequently asked how confident they are about the decision they just made. Metacognition is then a post-decision confidence judgment, i.e., the feeling that the prior decision was correct. Closer to economics, Enke and Graeber (2019) find that people vary regarding how uncertain they are about what the optimal action or solution to a decision problem is. They used the term "cognitive uncertainty" to design the amount of noise included in the decision process. In line with what is proposed in the present paper, Enke and Graeber (2019) try to link cognitive uncertainty to the ability to act rationally across *decision* problems. However, their paper differs from our study for two main reasons. First, they do not include games or strategic interactions in their decision problems. Second, they do not study to what extent those with high cognitive uncertainty would prefer not to perform the required task.

The transposition of a notion of *confidence in own performance* to strategic choices, such as non-cooperative games, is not straightforward. The feeling of having done (or not) the right thing when choosing a strategy in a game covers several dimensions. Players may not be confident in their understanding of the game, their ability to correctly map strategies to payoffs, or their ability to construct relevant beliefs about other players' behaviors. Rather than trying to single out one particular dimension, we here deliberately use a catch-all confidence judgment. The lack of precision associated with the multi-dimensional nature of confidence judgments in games is not a problem as long as we use confidence judgments to detect confused players (e.g., random players or players who play dominated strategies). This confusion arises when one or more dimensions of the decision process are impaired. Confusion and metacognition are thus likely to be intimately linked when it comes to games. Our claim is that players who feel confused, for whatever reason, will report low levels of confidence. Given the considerable fraction of subjects who appear

confused in one-shot games, the ability to identify confused players is important.[2]

Our decision to measure confusion through a confidence indicator was made for three reasons. First, subjects spontaneously express their feelings about the game in terms of confidence in informal debriefings at the end of experimental sessions. As such, the reporting of confidence seems natural for subjects. Second, confidence has proven to be a good predictor of actual performance in numerous experiments (although, to the best of our knowledge, the tasks involved were not games). Last, a direct measure of confusion needs to separate players who choose a strategy as a best response to their beliefs and those who randomly choose the same strategy. The identification of confused individuals, based only on their observed strategies, is therefore difficult.

Lastly, the use of confidence judgments may help contribute to the still open question regarding the stability of behaviors across games. A large within-subject variability in confidence judgment would suggest that a player who is confused in one game may not be so in a subsequent game. It is a proposition that is certainly worth testing regarding ongoing debates in behavioral game theory.

2.4. Promoting rationality

Standard economic theory is often considered as too demanding to describe the behavior of individuals. However, it has long been recognized that some mechanisms do promote rationality. We here review two mechanisms that have been studied in the lab since they may be field relevant: *learning* and *selection*. Studying learning is crucial since many economic transactions are taking place among participants who already learned through experience (List, 2003, 2004). Regarding selection, economic decision-makers may be different from the general population since they could have been selected according to different criteria (cognitive skills, experience) that may be linked to economic success. For instance, recruiters may choose candidates who obtained a specific diploma or scored well in a cognitive test. Furthermore, the market can select economic agents by eliminating participants who get bankrupt.

The *learning* mechanism (regarding strategic interactions) has been studied through repeated games. It has been shown that when agents receive feedback in repeated interactions, they converge to Nash equilibrium (see for instance Nagel, 1995; Weber et al., 2003; Gill and Prowse, 2016).

The *selection* mechanism has been explored by attempting to make lab subjects resemble decision-makers in the field and assessing whether they play closer to Nash equilibrium. Fréchette (2011) collected evidence from thirteen experiments involving professionals and

[2]Half a dozen strategies have been used to identify confusion. With different methodologies, all studies find a surprisingly-high fraction of confused subjects in one-shot games, of between 30% and 50% (see Costa-Gomes and Weizsäcker, 2008; Agranov et al., 2012; De Sousa et al., 2013; Burchardi and Penczynski, 2014; Agranov et al., 2015; Fragiadakis et al., 2016)

students. Apart from one experiment, professionals and students, facing games they are not familiar with, exhibit very similar behaviors. Levitt et al. (2010) also compares the behaviors of professional poker, bridge, and soccer players with students and do not find that professionals play closer to Nash in normal form games.

Our claim here is that a third force, self-selection, plays a role in filtering irrationality in many economic situations and thus must be investigated in the lab. We do not necessarily have insights into how the *self-selection* mechanism precisely combines with the two other forces at stake in the field, learning and selection but we expect self-selection to play a similar role.

3. Eliciting the determinants of self-selection in the lab: Experimental Design

Our experiment uses two well-known non-cooperative games: the undercutting and beauty-contest games (see below for details). In each game, after some understanding tests, subjects go through the following three stages. No feedback is given between stages, nor between games.

- Stage 1: Playing with all the subjects in the room.
- Stage 2: Choosing between a sure payoff and the one earned in Stage 1.
- Stage 3: Choosing between playing the game again (against those who decide to enter) and receiving a sure payoff

After playing all stages of the undercutting and beauty-contest games, two measures of risk-aversion and three measures of confidence are elicited.

A randomly-selected task is chosen for the payment. If a game is selected, the payoff is the average payoff against the strategies of all of the opponents in the group (i.e. the whole group in Stage 1, and the self-selected group in Stage 3). Those who chose the sure payoff receive the indicated amount if the corresponding task is selected. If one of the risk-aversion tasks is selected, the corresponding lottery is run.

The experiment was programmed using Z-Tree (Fischbacher, 2007).

3.1. Undercutting Game

Subjects are first asked to play the Undercutting Game shown in Figure 1.1; this is a symmetric normal-form game in which players have to *undercut* their opponent. The sure payoff in Stages 2 and 3 of this game is 15.

	1	2	3	4	5	6	7
1	16 16	5 25	15 15	15 15	15 15	15 15	15 4
2	25 5	15 15	5 25	15 15	15 15	15 15	15 15
3	15 15	25 5	15 15	5 25	15 15	15 15	15 15
4	15 15	15 15	25 5	15 15	5 25	5 25	5 25
5	15 15	15 15	15 15	25 5	15 15	15 15	15 15
6	15 15	15 15	15 15	25 5	15 15	15 15	15 15
7	4 15	15 15	15 15	25 5	15 15	15 15	4 4

Player 1 chooses a row and Player 2 a column.The bottom-left (respectively top-right) figure in the box corresponds to Player 1's (Player 2's) payoff. The payoffs are expressed in points.

Figure 1.1: Undercutting Game

The Undercutting Game possesses a unique Nash equilibrium: (1,1).

3.2. Beauty-Contest Game

Here subjects are asked to choose a whole number between 1 and 100. Their objective is to come as close as possible to the target of two-thirds of the average of all the numbers chosen by the other players. This avoids multiple equilibria (as both 0 and 1 are Nash equilibria when the set of available strategies is restricted to integers only). This game has a unique Nash equilibrium where all players choose 1. As compared to standard beauty-contest games, we here simplify the task by excluding the subject's own choice from the calculation of the mean. The payoffs are calculated using the following formula, which states that the payoff is proportional to the distance to the target:

$$\Pi(x_i) = 20 - \frac{1}{2}|x_i - \frac{2}{3(n-1)}\sum_{j \neq i} x_j| \tag{1.1}$$

The sure payoffs in Stages 2 and 3 for the beauty-contest game is 10. Sure payoffs are not the same in both games as we tried to calibrate them to be close to the empirical mean payoff in these games.

3.3. Risk-aversion

Players take the two risk-aversion elicitation tests introduced by Gneezy and Potters (1997) (GP) and Holt and Laury (2002) (HL)[3]. In the former they have to decide how much of a 10-token endowment they wish to invest in a risky asset. In the latter, they make a series of binary choices between lotteries. All subjects first take HL and then GP. These produce two measures of risk-aversion.

3.4. Confidence elicitation

At the end of the experiment, players state their feelings about the three tasks they carried out during the experiment: the first stages of the undercutting and beauty-contest games and the HL task. They are asked: *"When thinking about what to do in this task, you had:"*. The answers are on a 0 to 10 scale, with 0 meaning *"No idea of what to do"* and 10 *"A very clear idea of what to do"*.

3.5. Subjects

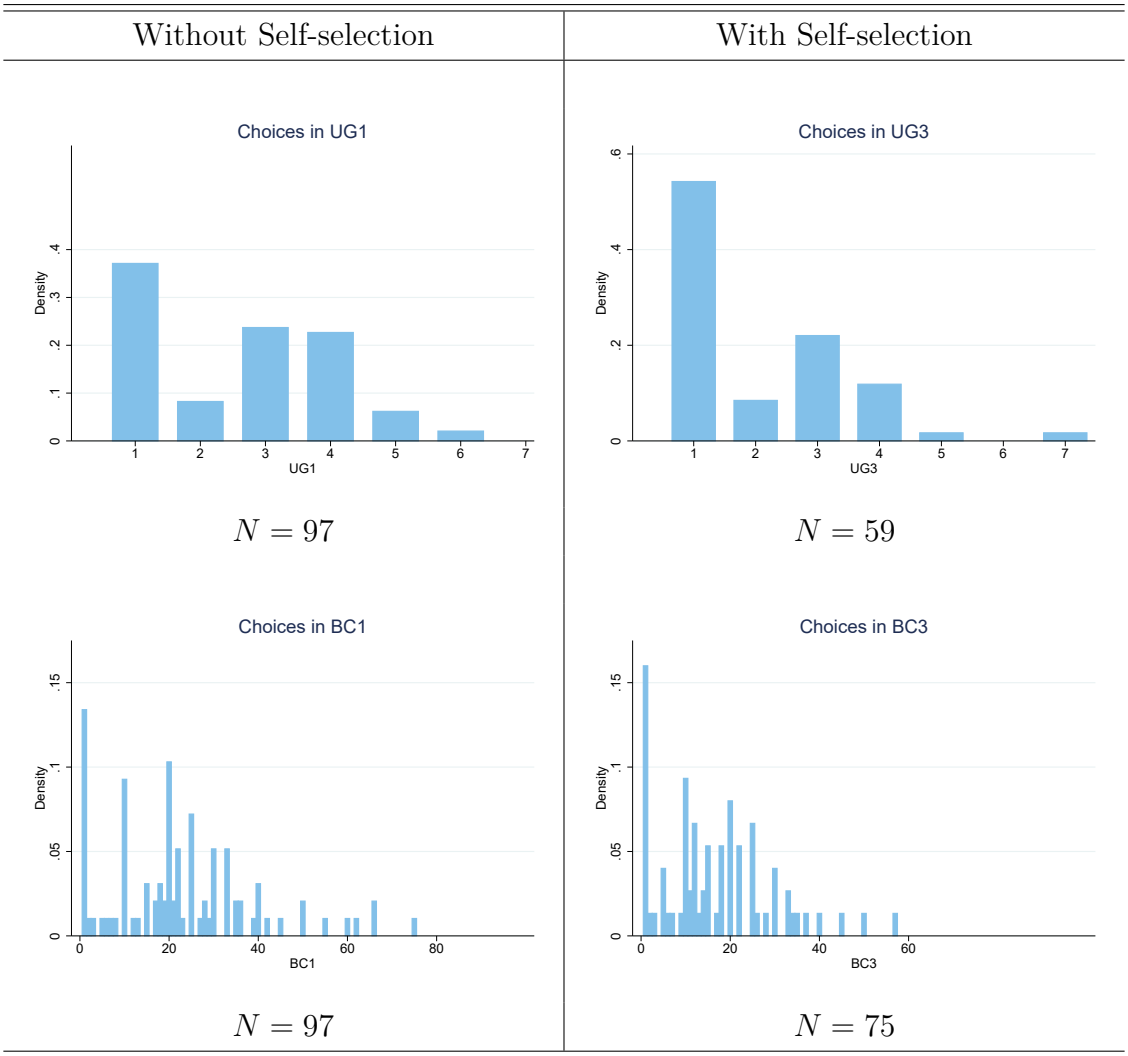
The experiment was carried out at the Ecole Polytechnique, which is widely-recognized as the top Engineering School in France. All subjects were students of the school, and so they are math savvy. The experiment lasted for about an hour. Four sessions were conducted with a total of 97 subjects (sample descriptive statistics appear in Table 1.9 in Appendix A). The average payoff was 16€.

4. Experimental Results

Figure 1.2 provides a first overview of the effect of self-selection. Self-selection does produce a substantial effect: we can see, for example, that it changes the mean strategy chosen from 22 to 14 in the beauty contest[4] and from 2.6 to 2.0 in the undercutting game. The numbers at the foot of each graph show that a substantial proportion of subjects prefer to opt out in Stage 3.

[3]We are aware that the uncertainty in games may more resemble ambiguous decisions (i.e. with unknown probabilities) than risky decisions with known probabilities. However, elicited measures of ambiguity-aversion are often noisy and inconsistent, even for subjects who are math-savvy. Furthermore, measures of risk- and ambiguity-aversion are often positively correlated (see Trautmann (2015)). We therefore limit ourselves to risk-aversion here.

[4]The figure of 22 is lower than that in comparable experiments, which typically produce average values of between 30 and 40. One explanation is that our subject pool is notable in terms of cognitive ability (Ecole Polytechnique students are strongly selected on the basis of their ability in Math and Science.)



Note: The blue bars show the fraction of active players who chose the given strategy. The number of active players appears at the foot of each graph.

Figure 1.2: Differences in strategies under self-selection

We first explore in more depth how strategies change with selection in the next subsection. We then look at the motivation for self-selection in a subsequent subsection. Last, we consider the link between strategic sophistication and metacognition.

4.1. How do strategies differ under self-selection?

To assess the evolution of strategies when adding a self-selection stage, we gather relevant variables in Table 1.2 and in Table 1.3.

Table 1.2: Undercutting Game

	Stage 1				Stage 3	
Strategy	Payoff	All	Out	In	Payoff	All
1	16.20	37%	32%	41%	16.21	54%
2	13.65	8%	13%	5%	11.72	8%
3	16.46	23%	18%	27%	15.17	22%
4	13.44	22%	32%	17%	13.10	12%
5	12.71	6%	3%	8%	13.79	2%
6	12.71	2%	3%	2%	13.81†	0%
7	12.51†	0%	0%	0%	13.79	2%
Mean	15.14	2.59	2.68	2.53	15.15	2.05
N^{o}. subjects	97	97	38	59	59	59

† indicates a hypothetical payoff against all subjects in the experiment, since that strategy was never chosen. *In* refers to those who subsequently self-select and *Out* to those who do not.

Table 1.3: Beauty-Contest Game

	Stage 1			Stage 3
	All	Out	In	All
Mean	22.46	32.09	19.64	16.49
Std Err	1.66	4.82	1.49	1.44
2/3 Mean	14.98	21.39	13.09	11.00
Payoff	13.23	8.52	14.61	14.91
Range		[1,34.89]		[1,30.87]
% in the range	81.44%	50%	90.67%	88%
N^{o}. subjects	97	22	75	75

The range indicates the set of strategies that provide a payoff greater than 10. *In* refers to those who subsequently self-select and *Out* to those who do not.

We list below six characteristics of the distribution of self-selected strategies, as compared to the non-selected ones.

1. **The strategies differ substantially under self-selection**. There is a clear and significant difference at all conventional levels between the distribution of strategies with and without self-selection, as revealed by a Mann-Whitney test.

2. **Entry is on average profitable**. Our design allows us to compare strategies with and without self-selection. For players who choose to enter, we can compute

actual payoffs under self-selection and compare these to the sure payoff offered as an alternative. Entry was profitable for 88% of players in the beauty contest and 76% in the undercutting game. We also note that the average payoff of those who did not self-select (calculated in Stage 1) is substantially lower. The average Stage-1 payoff in the beauty contest for those who subsequently self-select is 14.6 versus 8.5 for those who do not (with analogous figures of 15.3 vs. 14.9 in the undercutting game). In line with the existing literature, we thus find that decisions to self-select are rational on average.

3. **The fraction of individuals playing Nash increases**. There is a unique Nash equilibrium in the undercutting game, which is both players playing 1. The corresponding fraction of players who do so rose from 37% to 54% (i.e. +45%). In the beauty contest, the unique Nash equilibrium is all players choosing 1. This was initially chosen by 13% of subjects, which figure increased to 22% (+69%) with self-selection.

4. **There are fewer *poor* strategies** (i.e. either dominated or with low payoffs). In the undercutting game the proportion of dominated strategies (which are strategies 5, 6 and 7) fell from 8% to 4%. In the beauty contest, strategies over 66 are dominated. However, there are too few of these to make any meaningful comparisons. We can however compare the fraction of strategies of over 33, which correspond to level-0 players in the level-k model. About 17% of chosen strategies were strictly above 33 without self-selection, but only 6.5% with self-selection.

5. **Strategic sophistication is greater**. One common way of measuring strategic sophistication is to classify strategies as being at a particular level based on level-k models. We apply a charity principle, which assumes that selected strategies are assigned to the highest possible level. For instance if a level-3 strategy is chosen at random, it will be counted as level-3 (and not as level-0, as it should be). In the undercutting game, dominated strategies are considered as level-0, 4 corresponds to level-1, 3 to level-2, 2 to level-3, and 1 is Nash. Average strategic sophistication is higher under self-selection, with a first-order stochastic dominance of the distribution across levels. The same holds in the beauty contest: strategic sophistication rises under self-selection. There are several possible ways of assigning levels to players in the beauty contest. We here classify strategies over 34 as level-0, strategies in the range 32-34 as level-1, level-2 corresponds to any strategy between 21 and 31, and so on with each level-k best-responding to a group of players of level-$k-1$ and lower. Alternative classifications produce very similar results.

6. **Players do not adapt much to changes in the subject pool resulting from self-selection**. If players anticipate that self-selection will produce a more-strategic

subject pool, they should change their strategies. However, the strategies chosen do not differ much under self-selection. Players seem to underestimate the change in the subject pool: the second time they play, they will face more-strategic players and should choose more-sophisticated strategies, but do not do so. On average, those who entered neither learn[5] nor adapt much to self-selection, as can be seen in Tables 1.2 and 1.3 and Figure 1.3. In particular, Figure 1.3, shows that the distribution of strategies used in the beauty contest with self-selection (the green curve) is very similar to that in the first stage from the same players without self-selection (the blue curve). The red curve shows the strategies used by players who subsequently decided not to self-select. The same pattern holds in the undercutting game, but to a somewhat lesser extent.

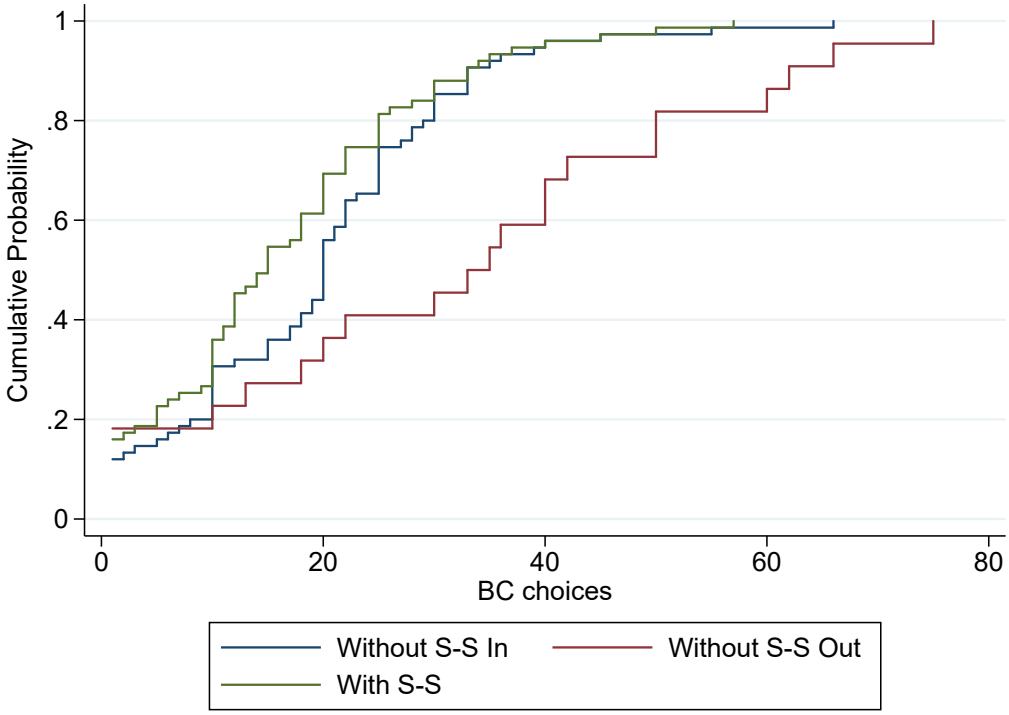


Figure 1.3: The Cumulative Distribution of Strategies in the Beauty Contest

We can summarize the six characteristics listed above by thinking of self-selection as a filter. The observed distribution of strategies changes as particular types of players (e.g., the most strategic) are selected; however, these self-selected players do not subsequently much adapt their strategies to the new subject pool.

[5]Even without feedback and self-selection, it has been shown that some kind of learning may nonetheless take place (see Weber et al., 2003, for evidence from repeated beauty-contest games without feedback).

4.2. Who self-selects?

We here consider two main sources of individual heterogeneity: the confidence measure (that is game-specific), and risk attitudes (more precisely, we look at the two measures of risk attitudes, as well as the combination of both measures[6]). The core of our analysis is probit regressions that estimate how much of the variance in entry decisions is captured by these two variables. We denote the entry decision of player k by y_k, with $y_k = 1$ if player k self-selects and $y_k = 0$ otherwise. For each probit regression, we calculate the fraction of concordant pairs. We look at all pairs (i, j) with $y_i = 1$ and $y_j = 0$, and declare them concordant if the predicted entry probability for i is greater than that for j. The results are shown at the foot of each column.

Table 1.4: Self-Selection: Undercutting Game

	(1)	(2)	(3)	(4)	(5)
Conf-UG	0.074***				0.077***
	(0.0190)				(0.0190)
Risk-GP		-0.021			
		(0.016)			
Risk-HL			0.030		
			(0.031)		
Risk-Comb.				-0.018	-0.056
				(0.0673)	(0.0617)
N	97	97	97	97	97
Concordant Pairs	70.2%	56.0%	53.8%	51.3%	71.3%

The figures here are the average marginal effects and robust standard errors (in parentheses). * p<0.1, ** p<0.05, *** p<0.01.

[6] Gillen et al. (2019) and Perez et al. (2021) show the existence of measurement error in risk-aversion measures, and evaluate its consequences. Gillen et al. (2019) suggest to elicit two measures and to use one as an instrument of the other measure. This IV strategy is of particular relevance when both measures are sufficiently correlated. As it is not the case here, we created a risk-aversion indicator, which roughly corresponds to the average of the two measures. Details appear in the Appendix B.

Table 1.5: Self-Selection: Beauty Contest

	(1)	(2)	(3)	(4)	(5)
Conf-BC	0.032*				0.023
	(0.019)				(0.018)
Risk-GP		-0.041***			
		(0.0117)			
Risk-HL			-0.036		
			(0.032)		
Risk-Comb.				-0.17***	-0.16***
				(0.049)	(0.049)
N	97	97	97	97	97
Concordant Pairs	58.2%	71.2%	63.7%	70.4%	70.1%

The figures here are the average marginal effects and robust standard errors (in parentheses). * p<0.1, ** p<0.05, *** p<0.01.

The probit estimates in Table 1.4 and Table 1.5 suggest that our two variables of interest account for a substantial part of the decision to self-select into games. In the undercutting game, 71% of the pairs are concordant (this figure would be 50% were the entry decisions to be randomly-chosen, and 100% if the model was perfect) and in the beauty-contest game this figure is 70%. It is notable that the respective contributions of confidence and risk attitude vary between games. Confidence is the main driver of self-selection in the undercutting game, while risk attitude is more important in the beauty contest. Table 1.6 shows the correlations between the explanatory variables in Table 1.4 and Table 1.5, and underlines that confidence and risk attitudes seem to capture different dimensions of individual behavior, as they are not correlated with each other.

Table 1.6: Correlations among explanatory variables

	Conf-UG	Conf-BC	Risk-GP	Risk-HL	Risk-Comb.
Conf-UG	1.00				
Conf-BC	0.08	1.00			
Risk-GP	0.09	-0.12	1.00		
Risk-HL	0.14	-0.10	0.11	1.00	
Risk-Comb.	0.16	-0.15	0.75***	0.75***	1.00

These are the correlation coefficients between the explanatory variables. The correlations between Risk-Comb. and Risk-HL/RiskGP are mechanical, from the construction of the former.

4.3. Strategic sophistication and Metacognition

We suggest that players in experimental games have a sense of how well they are doing and that a proxy of this feeling can be observed by the experimenter. We here confront that assumption to our data. We already provided evidence that the most confused players are less likely to self-select in the previous subsection. Two new pieces of evidence can be added.

We can first test whether elicited levels of confidence are linked to chosen strategies in games in which all players participate (stage 1 in our design). As a first approximation, we can look at the correlation between confidence measures and chosen strategies. Indeed, in both games choosing low numbers can be considered as a measure of strategic sophistication; in particular because both the beauty contest game and the undercutting game are dominance solvable, with iterative elimination of dominated strategies excluding the highest numbers first. Therefore, a crude assessment of the subject's awareness of their strategic sophistication is to look at the correlations between their choices in the game and their confidence indicator for each game. A second test on the ability of subjects to evaluate their own (relative) ability to play games is to check that confidence in a specific game (for instance the beauty context game in our protocol) is less predictive that confidence in another game (for instance the undercutting game). Table 1.7 and Table 1.8 show all correlation estimates between confidence indicators and strategies and thus allow us to draw conclusions on the assumptions we wanted to test.

Table 1.7: Empirical correlations between strategies and confidence indicators

	UG1 Strat.	BC1 Strat.	Conf-UG	Conf-BC
UG1 Strat.	1.00			
BC1 Strat.	0.08	1.00		
Conf-UG	-0.31***	-0.04	1.00	
Conf-BC	-0.03	-0.42***	0.08	1.00

These are the empirical correlation estimates between chosen strategies and confidence indicators.* p<0.1, ** p<0.05, *** p<0.01.

Table 1.8: Spearman Correlation between strategies and confidence indicators

	UG1 Strat.	BC1 Strat.	Conf-UG	Conf-BC
UG1 Strat.	1.00			
BC1 Strat.	0.16	1.00		
Conf-UG	-0.31***	-0.08	1.00	
Conf-BC	-0.03	-0.45***	0.16	1.00

These are the Spearman correlation estimates between chosen strategies and confidence indicators.* p<0.1, ** p<0.05, *** p<0.01..

1. Subjects are, on average aware of their strategic sophistication. Indeed, the BC confidence measure strongly correlates negatively with the number chosen in the first BC game. The UG confidence measure also strongly negatively correlates with the number chosen in the first UG.

2. Confidence measures are game-specific. The BC (resp UG) confidence measure does not correlate with the strategies in the first UG (resp BC). Both confidence measures are not significantly correlated. Those simple non-incentivized confidence tasks seem to isolate well a game-specific component related to strategic sophistication or a sense of a clear idea of what to do in a specific game.

In sum, economic agents can gauge how well they are doing in a specific experimental game, even without any feedback and may choose to enter or not in an experimental games based on their feelings.

On the stability of types

The question of whether individual traits regarding behavior in games are stable is a rather open one. For example, IQ is typically found to increase strategic sophistication, but with only small effect sizes. Georganas et al. (2015) explore in a systematic way whether types (as defined in level-k models) are stable across games, and find only little stability if any. Our confidence measure captures an important aspect of strategic behavior, but observation 2 above highlights the context-dependence of confidence. Therefore, stability of types across strategic environments might be compromised by this instability of confidence.

Futhermore, We note that our subjects are at the very top of the distribution in terms of mathematical ability. We do not have a direct measure of cognitive skills for the subjects in our experiment. However, using a different group of students at the same institution (Ecole Polytechnique), we found that almost all completed the Cognitive Reflection Test (CRT, see Frederick, 2005) without error[7], suggesting that confidence is not just a proxy for cognitive skills, but rather captures a separate dimension that is game-specific.

[7]The mean score of a group of about 100 subjects was 2.67, where 3 indicates a perfect score.

5. Self-Selection : Evidence from the field

What we are after in the present section is field evidence on (1) whether agents have the possibility to self-select into strategic environments, (2) whether strategies under self-selection are different than strategies when there is a compulsory participation (in particular whether strategies under self-selection are closer to theoretical predictions and somehow more homogeneous) and (3) whether agents that self-selected into strategic interactions have different characteristics such as risk-attitude (as we find in our lab experiment).

To the best of our knowledge, it is difficult to collect field data that allows external observers to compare behavior with and without self-selection. There are not many situations in which economic agents may be forced to participate. Furthermore, we need to assess what rationality means, which implies that something like an optimal strategy can be defined. Primitives of game theory (an explicit link from chosen strategies to payoff in particular) are usually difficult to observe in practice. As a result, field tests of game theory usually come from TV shows, sports, elections and auctions. So we carefully review corresponding literature for information regarding how the behavior of self-selected subject pools can be compared to that of non-selected ones and find relevant examples related to games in the field, to voting or to auctions.

5.1. Games in the field

An interesting instance of games played in the field is the series of beauty contest experiments reported in Bosch-Domenech et al. (2002). Authors present results from a series of beauty contests so that, as in our experiments, subjects have the opportunity to self-select into the task *once they learned about it*. Readers of three newspapers were offered the possibility to submit an answer to a beauty-contest game. Interestingly, participating implies to pay the cost of sending a physical letter (the experiment took place before the Internet) and winners got substantial prizes. Not participating entails a sure outcome with neither cost nor gain. This is exactly the type of self-selection into a given task we are interested in. A perfect counterfactual would have been to have the *same* pool of readers being forced to participate. Unfortunately, such data do not exist. However, having subjects joining a lab experiment, and subsequently offered to play the beauty contest, allows a crude comparison of self-selected subjects vs non self-selected ones. We note that such a comparison is certainly imperfect as we ignore self-selection into the lab, so as to treat observations as resulting form a subject pool composed of a random sample of population similar to the readers of the considered newspapers. Self-selected subject pools indeed are using strategies that are closer to Nash equilibrium and dominated strategies almost disappeared. Östling et al. (2011) analyse another game that has been played in

the field: the LUPI (lowest unique positive integer) Swedish lottery. Players self-select into this game since there is a fix cost for participating. Unfortunately, we do not have a counterfactual in which players are forced to play the same game[8]. However, it is worth noticing that Östling et al. (2011) find that players strategies are not far from theoretical predictions in this setting in which players self-select.

5.2. Voting

Elections are compulsory in some countries but not in other ones. In large elections, there is little room for strategic voting, so the most relevant strategy would be for each voter to support the candidate that best fits her political views. A natural test of our assumption is thus to check whether there is a better match between voter's ballot choices and preferences among those who would have not voted if voting was not compulsory. There is strong evidence supporting this view (Singh and Roy, 2018). Furthermore, voting appears more dispersed in countries with compulsory voting, with more vote for extreme parties, compatible with more *random* voting. We find, there too, evidence that self-selection leads to more homogeneous groups and lower levels of irrationality. Another interesting observation is the fact that when ask to participate to multiple ballots on the same day, at the same place, some voters choose to vote for some ballots but not for others, as if they self-select into the ballots they have a clear view on.

5.3. Auctions

Auctions are strategic interactions in which individuals are free to participate. However, identification of utility functions based on bidding behavior has proven difficult and thus requires some assumptions. A recent attempt is a study by Grundl and Zhu (2019) who estimates a structural model of bidding in auctions using timber auctions in the US. Although their specification does allow for non-risk neutral bidders, they cannot reject the assumption that active bidders are risk-neutral while the general population is known to be risk-averse (see for instance Dohmen et al. (2005)). Ackerberg et al. (2006) also provide interesting evidence that when offered the option to either buy an object at a fixed price or enter in a competitive auction, a substantial fraction of eBay customers prefers to buy at a fixed (and higher) price rather than entering into a strategic interaction. This result is not straightforward to interpret. We suggest that it proves that risk-averse bidders would prefer a sure payoff. In any case, this opt-out option acts as a sure payoff alternative. Some agents prefer to shy away from strategic interactions, comforting our idea of self-selection in many economic situations.

[8]The LUPI game has been played in the lab but the rules were different and more importantly, the number of players was much lower, a parameter that has a tremendous impact on strategies

6. Self-selection and external validity

In the present short section, we discuss why we believe that adding an explicit self-selection stage in the lab can enhance the external validity of experiments.

As stated by List (2020) "In its simplest form, questions of external validity revolve around whether the results of the received study can be generalized to different people, situations, stimuli, and time periods." List then outlines four key areas that every study should report to address external validity, called the SANS conditions.[9] We argue that introducing a self-selection stage increases external validity in two areas listed by List.

Selection of subjects: usually, subject pools are deemed comparable if they share common characteristics. We here suggest that self-selection makes subject pools more comparable because self-selected subjects share important characteristics (e.g. risk attitude, understanding of strategic interactions) that are key determinants when it comes to choose a strategy in one-shot games. Observing the behavior of self-selected subjects in a particular one-shot game helps predicting behavior of different subject pools playing the same game in different contexts.

Naturalness: The *Naturalness* condition tackles among other things the adequacy between lab and field settings. Since in many economic situations (auctions, market, elections), participation is not mandatory, strategic interactions mostly take place among self-selected subject pools. Adding a self-selection stage in lab and field experiments thus appears natural.

In brief, we think that games played in the lab are more predictive of field situations when introducing a self-selection stage. We hope to open a discussion on how self-selection stages could improve more generally the predicting power of lab experiments.

7. Discussion

Choosing between A and B may not be straightforward if A is receiving a sure payoff and B is entering a strategic interaction. Indeed, potential players need to evaluate future payoffs resulting from strategic interactions. However, anticipating future payoffs in a game is not an easy task, particularly in situations where little learning has occurred. Our main finding is that economic agents are endowed with a feeling of how well they are doing in a given game and that this feeling greatly varies across games. The main consequence we here focus on is linked to self-selection. On average, agents can make sensible entry decisions, avoiding situations in which they have low confidence and privileging environments where they feel comfortable. Self-selection plays thus a similar role as learning: it prevents the most blatant mistakes and homogenizes the pool of interacting agents. We note that the main effect is filtering, much more than adaptation to changes

[9]The four SANS conditions are Selection of subjects, Attrition, Naturalness and Scaling

in characteristics of the subject pool. The behavior of self-selected pools is thus easier to predict and closer to the Nash benchmark. Self-selection reduces the gap between actual plays in one-shot games and Nash equilibrium.

At a more general level, we here contribute to the growing literature linking confidence measures, or cognitive uncertainty, to observed deviations away from rational decisions. To the best of our knowledge, we here provide a first step regarding how games, rather than individual decisions, are linked to confidence. We here focus on self-selection. We find that agents have a feeling of how well they are doing, which is context-specific. At the same time, they are much less accurate in considering the impact of self-selection on the composition of the pool of players. We believe that these findings contribute to a better understanding that the ability to play a game is context-specific, i.e., the fact that experts in one field (like poker or chess players) do not seem to perform any better in a different context. We also contribute to the literature showing that strategic sophistication varies substantially across games.

References

Ackerberg, D., K. Hirano, and Q. Shahriar (2006). The buy-it-now option, risk aversion, and impatience in an empirical model of ebay bidding. *Unpublished Manuscript, University of Arizona*.

Agranov, M., A. Caplin, and C. Tergiman (2015). Naive play and the process of choice in guessing games. *Journal of the Economic Science Association 1*(2), 146–157.

Agranov, M., E. Potamites, A. Schotter, and C. Tergiman (2012). Beliefs and endogenous cognitive levels: An experimental study. *Games and Economic Behavior 75*(2), 449–463.

Ais, J., A. Zylberberg, P. Barttfeld, and M. Sigman (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition 146*, 377–386.

Bosch-Domenech, A., J. G. Montalvo, R. Nagel, and A. Satorra (2002). One, two,(three), infinity,...: Newspaper and lab beauty-contest experiments. *American Economic Review 92*(5), 1687–1701.

Brekke, K. A., K. E. Hauge, J. T. Lind, and K. Nyborg (2011). Playing with the good guys. a public good game with endogenous group formation. *Journal of Public Economics 95*(9-10), 1111–1118.

Burchardi, K. B. and S. P. Penczynski (2014). Out of your mind: Eliciting individual reasoning in one shot games. *Games and Economic Behavior 84*, 39–57.

Camerer, C. and D. Lovallo (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review 89*(1), 306–318.

Choo, L., T. R. Kaplan, and X. Zhou (2019). Can auctions select people by their level-k types? *Available at SSRN 3451312*.

Choo, L. and X. Zhou (2019). Can market competition reduce anomalous behaviours. Technical report, FAU Discussion Papers in Economics.

Cooper, D. J., C. A. Ioannou, and S. Qi (2018). Endogenous incentive contracts and efficient coordination. *Games and Economic Behavior 112*, 78–97.

Costa-Gomes, M. A. and G. Weizsäcker (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies 75*(3), 729–762.

Dana, J., D. M. Cain, and R. M. Dawes (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and human decision Processes 100*(2), 193–201.

De Sousa, J., G. Hollard, and A. Terracol (2013). Non-strategic players are the rule rather than the exception. Technical report, working paper, Université Paris 1.

Dohmen, T. and A. Falk (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review 101*(2), 556–90.

Dohmen, T. J., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2005). Individual risk attitudes: New evidence from a large, representative, experimentally-validated survey.

Enke, B. and T. Graeber (2019). Cognitive uncertainty. Technical report, National Bureau of Economic Research.

Eriksson, T., S. Teyssier, and M.-C. Villeval (2009). Self-selection and the efficiency of tournaments. *Economic Inquiry 47*(3), 530–548.

Exley, C. L., M. Niederle, and L. Vesterlund (2020). Knowing when to ask: The cost of leaning in. *Journal of Political Economy 128*(3), 816–854.

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

Fleming, S. M. and H. C. Lau (2014). How to measure metacognition. *Frontiers in Human Neuroscience 8*, 443.

Fragiadakis, D. E., D. T. Knoepfle, and M. Niederle (2016). Who is strategic? *Unpublished manuscript, Stanford University, Stanford, CA*.

Fréchette, G. R. (2011). Laboratory experiments: Professionals versus students. *Available at SSRN 1939219*.

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

Georganas, S., P. J. Healy, and R. A. Weber (2015). On the persistence of strategic sophistication. *Journal of Economic Theory 159*, 369–400.

Gill, D. and V. Prowse (2016). Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis. *Journal of Political Economy 124*(6), 1619–1676.

Gillen, B., E. Snowberg, and L. Yariv (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy 127*(4), 1826–1863.

Gneezy, U. and J. Potters (1997). An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics 112*(2), 631–645.

Grundl, S. and Y. Zhu (2019). Identification and estimation of risk aversion in first-price auctions with unobserved auction heterogeneity. *Journal of Econometrics 210*(2), 363–378.

Heckman, J. J. (1990). Selection bias and self-selection. In *Econometrics*, pp. 201–224. Springer.

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92*(5), 1644–1655.

Lazear, E. P., U. Malmendier, and R. A. Weber (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics 4*(1), 136–63.

Levitt, S. D., J. A. List, and D. H. Reiley (2010). What happens in the field stays in the field: Exploring whether professionals play minimax in laboratory experiments. *Econometrica 78*(4), 1413–1434.

List, J. A. (2003). Does market experience eliminate market anomalies? *The Quarterly Journal of Economics 118*(1), 41–71.

List, J. A. (2004). Neoclassical theory versus prospect theory: Evidence from the marketplace. *Econometrica 72*(2), 615–625.

List, J. A. (2020). Non est disputandum de generalizability? a glimpse into the external validity trial. Technical report, National Bureau of Economic Research.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review 85*(5), 1313–1326.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The quarterly journal of economics 122*(3), 1067–1101.

Östling, R., J. T.-y. Wang, E. Y. Chou, and C. F. Camerer (2011). Testing game theory in the field: Swedish lupi lottery games. *American Economic Journal: Microeconomics 3*(3), 1–33.

Perez, F., G. Hollard, and R. Vranceanu (2021). How serious is the measurement-error problem in risk-aversion tasks? *Journal of Risk and Uncertainty*, 319–342.

Shachat, J. and J. T. Swarthout (2013). Auctioning the right to play ultimatum games and the impact on equilibrium selection. *Games 4*(4), 738–753.

Singh, S. P. and J. Roy (2018). Compulsory voting and voter information seeking. *Research & Politics 5*(1), 2053168017751993.

Trautmann, S. T. (2015). Ambiguity attitudes. *The Wiley Blackwell Handbook of Judgment and Decision Making 1*, 89–116.

Weber, R. A. et al. (2003). 'Learning' with no feedback in a competitive guessing game. *Games and Economic Behavior 44*(1), 134–144.

Appendix 1.A Subjects

Table 1.9: Descriptive Statistics regarding the subject pool

	Gender		Age		Background		Participation	Game theory
	Number	(%)	Mean	SD	Scientific	Other	Yes (%)	Yes (%)
Male	77	79	20.40	0.976	99	1	9	26
Female	20	21	20.45	1.276	100	0	4	5
All	97	100	20.41	1.038	99	1	8	22

Participation refers to a previous participation organized by the Labeds. Game Theory refers to the fact that the student has followed a course of game theory

Appendix 1.B Descriptive Statistics

Tasks 1 to 3: Undercutting Game

We here show, for each stage, the distribution of strategies and the corresponding payoffs. We also show the distribution of Stage-1 strategies for two sub-groups: those who will self-select in Stage 3 and those who will not. We can see a clear difference between the strategies that the two groups use.

Table 1.10: Undercutting Game

	Stage 1						Stage 3	
			Choices Stage 2		Choices Stage 3			
Strategy	Payoff	All	Out	In	Out	In	Payoff	All
1	16.20	37%	28%	44%	32%	41%	16.21	54%
2	13.65	8%	13%	5%	13%	5%	11.72	8%
3	16.46	23%	20%	26%	18%	27%	15.17	22%
4	13.44	22%	25%	21%	32%	17%	13.10	12%
5	12.71	6%	10%	4%	3%	8%	13.79	2%
6	12.71	2%	5%	0%	3%	2%	13.81†	0%
7	12.51†	0%	0%	0%	0%	0%	13.79	2%
Mean	15.14	2.59	2.93	2.35	2.68	2.53	15.15	2.05
N^{o}. subjects	97	97	40	57	38	59	59	59

† indicates a hypothetical payoff against all subjects in the experiment since that strategy was never chosen.

Tasks 4 to 6: Beauty-Contest Game

We here provide some basic information on the distribution of strategies (means and standard errors) in the beauty-contest game. We also list these figures according to whether the subject subsequently self-selects in Stage 3. As for the undercutting game, we find a difference between these two distributions.

Table 1.11: Beauty-Contest Game

	Stage 1					Stage 3
		Choices Stage 2		Choices Stage 3		
	All	Out	In	Out	In	All
Mean	22.46	30.16	19.79	32.09	19.64	16.49
Std Err	1.66	4.30	1.57	4.82	1.49	1.44
2/3 Mean	14.98	20.11	13.19	21.39	13.09	11.00
Payoff	13.23	9.63	14.47	8.52	14.61	14.91
Range			[1,34.89]			[1,30.87]
% in the range	81.44%	60.00%	88.89%	50%	90.67%	88%
N°. subjects	97	25	72	22	75	75

The range indicates the set of strategies that provide a payoff larger than the proposed sure payoff (10). % in range indicates the proportion of strategies that yield a payoff inside the range.

Tasks 7 to 8: Risk-Aversion Measures

Our protocol includes two measures of risk: Gneezy-Potter (GP henceforth) and Holt-Laury (HL). The number of safe choices in the Gneezy-Potter task corresponds to the number of tokens that are retained; in the HL task this corresponds to the number of safe choices (i.e. the number of choices in the left-hand column, which yields a payoff of between 16 and 20, while this varies from 1 to 39 for choices in the right-hand column). Table 1.12 presents the frequency of each choice in the two tasks. In the GP task 43.3% of subjects invest all of their endowment, suggesting that they are either risk-neutral or risk-loving. On the contrary, the HL task indicates that the subject pool is mostly on the risk-averse side (corresponding to five or more safe choices). The large fraction of subjects choosing 4 can be interpreted as expected-gain maximizers. This is consistent with the large fraction of subjects choosing to invest all of their tokens in the Gneezy-Potter task.

Table 1.12: Gneezy-Potters and HL summary

Tokens Kept	Freq.	Percent	Safe Choices HL	Freq.	Percent
0	42	43.30	0	1	1.03
1	1	1.03	1	0	0.00
2	6	6.19	2	0	0.00
3	4	4.12	3	3	3.09
4	8	8.25	4	30	30.93
5	16	16.49	5	14	14.43
6	7	7.22	6	23	23.71
7	5	5.15	7	16	16.49
8	4	4.12	8	8	8.25
9	1	1.03	9	2	2.06
10	3	3.09	10	0	0.00
Total	97	100.0	Total	97	100.00

The two risk measures are poorly-correlated, suggesting large within-subject variations. We combine the two measures. We re-scale them to have zero mean and a variance of 1, and calculate the mean of the two re-scaled measures. The corresponding distribution appears in Figure 1.4.

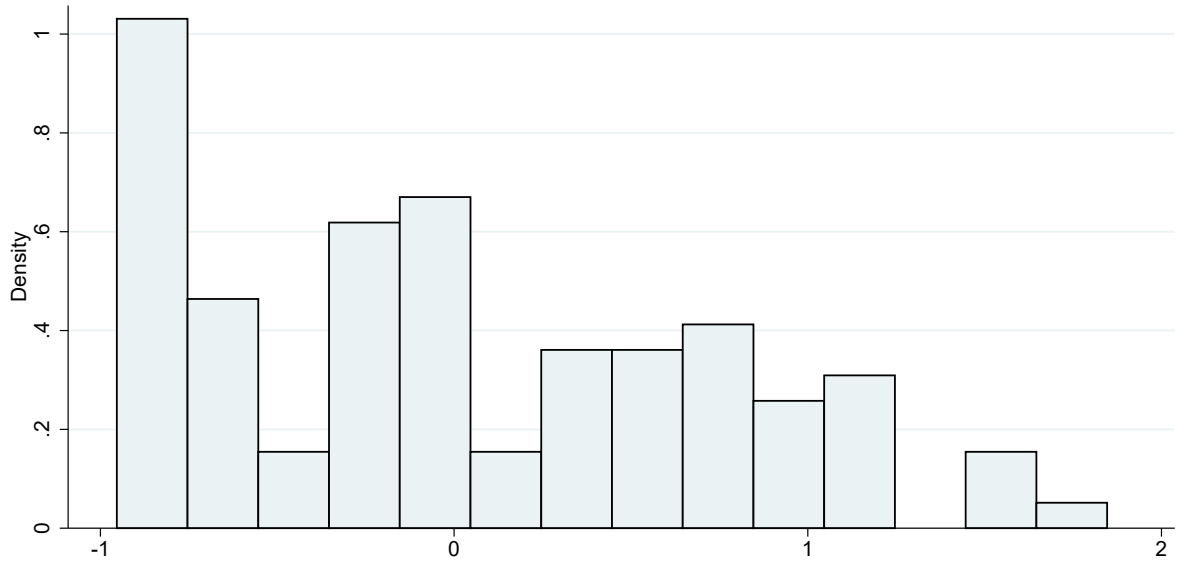


Figure 1.4: The Constructed Risk-Aversion Indicator

Task 9: Confidence Task

Table 1.13 indicates the distribution of confidence measures in each game. Confidence seems to be heterogeneous across individuals. Few individuals use the lower end of the scale with values of 0, 1, 2 and 3. About 75% report confidence at the top end of the

scale, starting at 7.

Table 1.13: Confidence Measure

	Undercutting Game		Beauty-Contest Game	
Confidence	Freq.	Percent	Freq.	Percent
0	0	0.00	0	0.00
1	1	1.03	0	0.00
2	2	2.06	3	3.09
3	2	2.06	3	3.09
4	9	9.28	4	4.12
5	5	5.15	6	6.19
6	8	8.25	8	8.25
7	12	12.37	15	15.46
8	31	31.96	18	18.56
9	9	9.28	19	19.59
10	18	18.56	21	21.65
Total	97	100	97	100

Appendix 1.C Probit Estimates for Stage 2

Table 1.14: Choosing past earnings over the sure payoff: Undercutting Game

	(1)	(2)	(3)	(4)	(5)
Confidence	0.10***				0.11***
	(0.016)				(0.015)
Risk-GP		0.0040			
		(0.017)			
Risk-HL			-0.045		
			(0.034)		
Risk-Comb.				-0.055	-0.11**
				(0.066)	(0.055)
N	97	97	97	97	97
Concordant Pairs	76.6%	52.1%	60.2%	54.6%	78.2%

The figures here are the average marginal effects and robust standard errors (in parentheses). * p<0.1, ** p<0.05, *** p<0.01.

Table 1.15: Choosing past earnings over the sure payoff: Beauty Contest

	(1)	(2)	(3)	(4)	(5)
Confidence	0.010				0.00055
	(0.021)				(0.0206)
Risk-GP		-0.024*			
		(0.0136)			
Risk-HL			-0.066**		
			(0.027)		
Risk Comb				-0.15***	-0.15***
				(0.050)	(0.050)
N	97	97	97	97	97
Concordant Pairs	52.9%	61.6%	66.1%	67.9%	67.7%

The figures here are the average marginal effects and robust standard errors (in parentheses). * p<0.1, ** p<0.05, *** p<0.01.

Appendix 1.D Instructions

(Available as a printed copy for each subject)

Introduction

Welcome to our experiment and thank you for participating!
Please read the instructions carefully.

The amount you will earn at the end of the experiment will depend on your decisions, the decisions of the other participants and chance. In addition, you will receive a participation fee of 7.5€. Your earnings will be paid individually and in cash immediately at the end of the experiment; no other participant will know how much you earned.

All amounts in the experiment will be expressed in points. At the end of the experiment, the points you earned will be converted into Euros using the following exchange rate:

2 points = 1€

The participation fee of 7.5€ therefore corresponds to 15 points.

During this experiment, you will face nine tasks. In each task, you can either earn or lose points. At the end of the experiment, one task will be randomly selected. You will be paid according to the points you earned in that task. For instance, imagine that your number of points for each task is as follows:

- Task 1: 8 points
- Task 2: 12 points
- Task 3: -4 points
- Task 4: 6 points
- Task 5: 18 points
- Task 6: -1 point
- Task 7: 4 points
- Task 8: 6 points
- Task 9: 20 points

In this example, if Task 4 is randomly selected, you would earn the participation fee (15 points) plus 6 points, that is 21 points (10.5€). If Task 6 is randomly selected, you will earn $15 - 1 = 14$ points (7€). You can see that it is very important that you do your best in each and every task.

You will make your decisions by clicking on the appropriate buttons on the screen or typing answers on the keyboard. All participants read the same instructions and are taking part in this experiment for the first time, as you are.

Please note that hereafter any form of communication between participants is strictly prohibited. If you violate this rule, you will be excluded from the experiment with no payment. If you have any questions, please raise your hand. The experimenter will come to you and answer your questions individually.

Instructions for Tasks 1 to 3

In the first three tasks your gains will depend on your decisions as well as those of other participants.

In these three tasks you may face some games. For each task, the instructions will be displayed on your screen at all times. Nonetheless, for you to fully understand how your gains will be calculated, we set out here how the games are played and then ask you to take a small understanding test.

You \ Other	1	2	3	4
1	14 / 14	-34 / -10	18 / 26	8 / 19
2	-10 / -34	24 / 24	14 / 13	-16 / 16
3	26 / 18	13 / 14	-2 / -2	12 / 53
4	19 / 8	16 / -16	53 / 12	-45 / -45

Figure 1.5: Game Example 1

Game Example 1 above is an example of the games you will face. In this game, you have to choose a row, corresponding to a number (**in black**) in the left column. Your opponent (another player from the room) will choose a column, corresponding to a number (**in blue**). After you and your opponent have chosen your strategies, your points will be the number **in black** (at the bottom left-hand corner) in the box at the intersection of the row you have chosen and the column your opponent chooses. The points of your opponent are colored **in blue** (in the top-right hand corner) in the same box.

For instance, in Game Example 1, if you choose 3 and your opponent chooses 4, you will earn 53 points and your opponent will earn 12. If you choose 1 and your opponent chooses 2, you will earn -10 (lose 10 points) and your opponent will earn -34 (lose 34 points).

Important detail regarding the calculation of your payoff

In each game you play, your choice will be matched to each choice made by all of the other participants in the experiment. Your earnings in points will then be the average earnings you would receive from playing individually against all of the other participants.

Appendix 1.E Experiment Screenshots

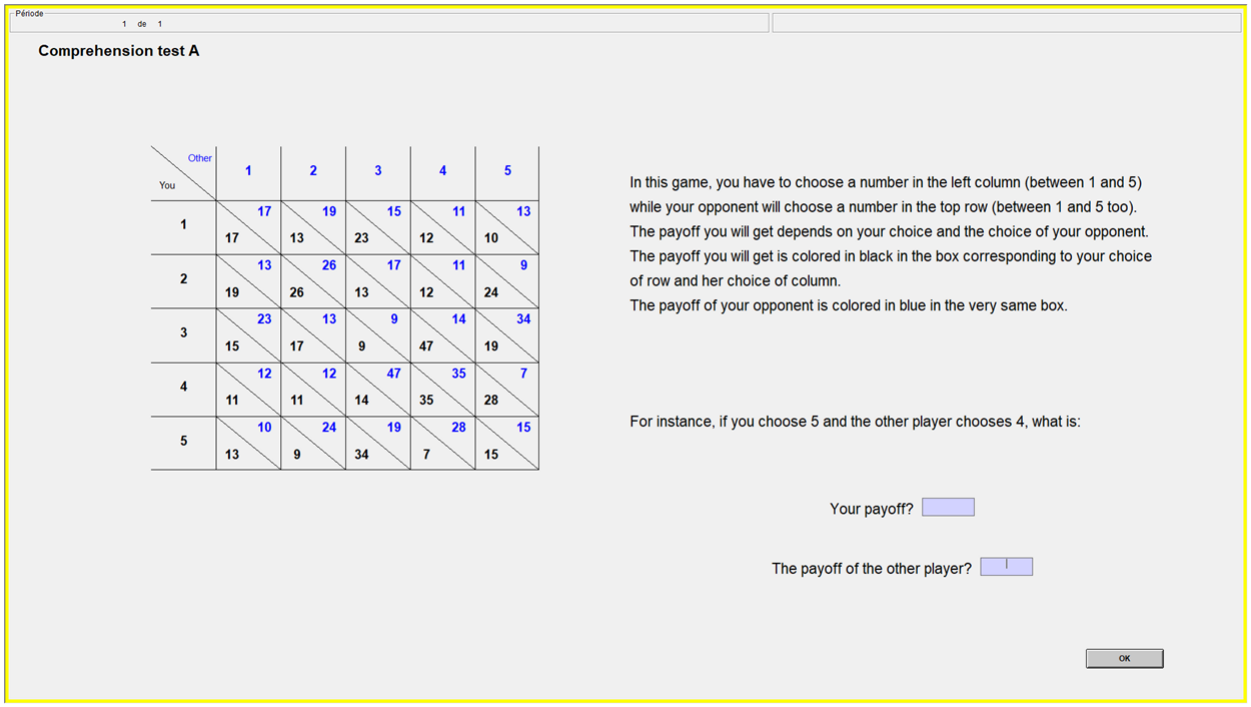


Figure 1.6: Trial 1

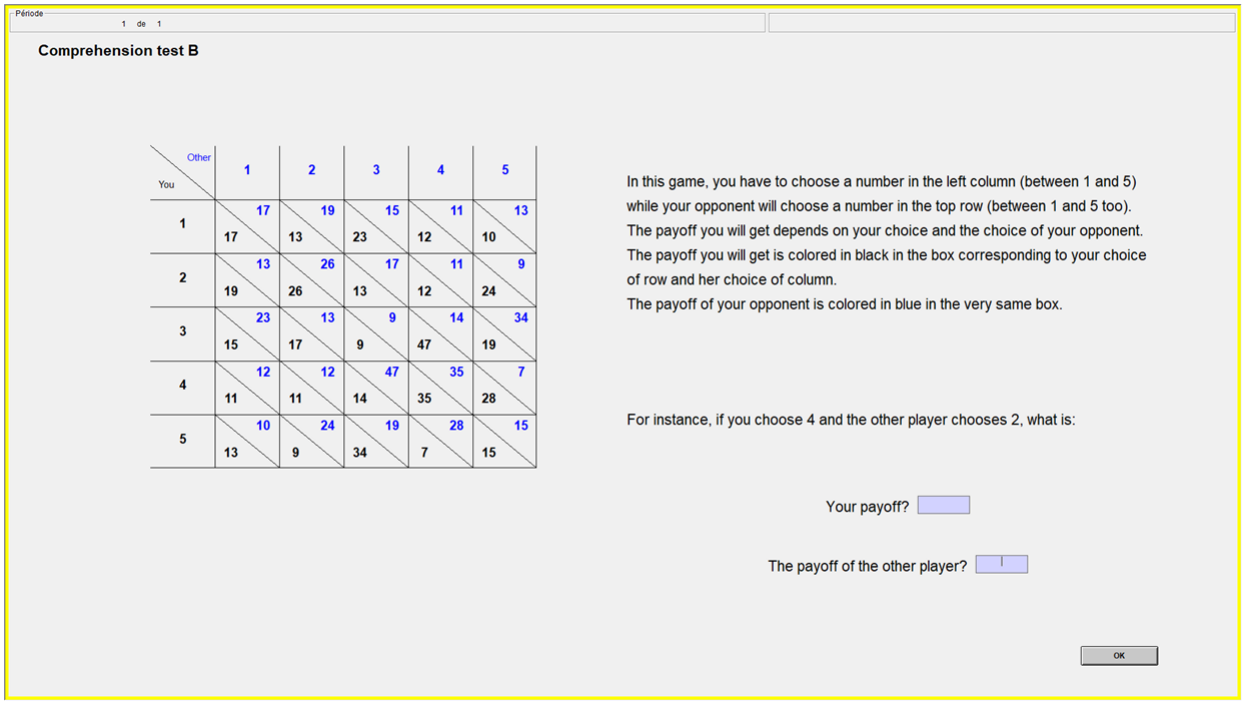


Figure 1.7: Trial 2

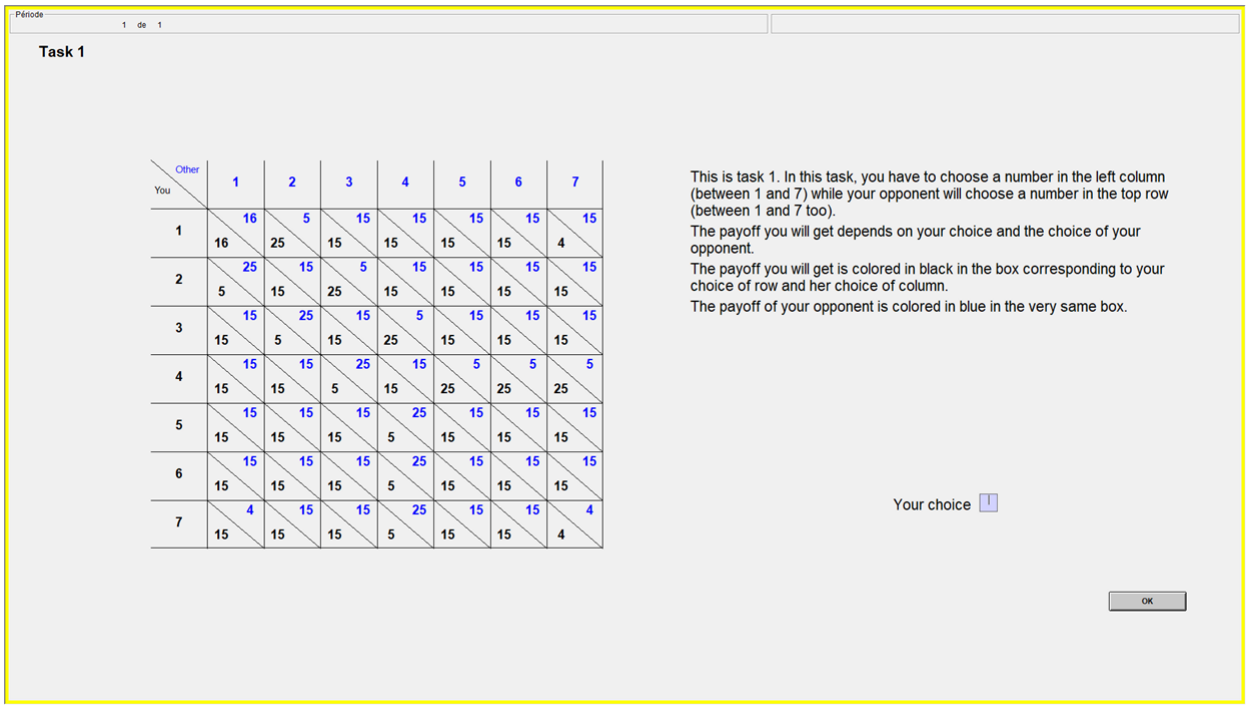


Figure 1.8: Task 1 (Undercutting Game: Stage 1)

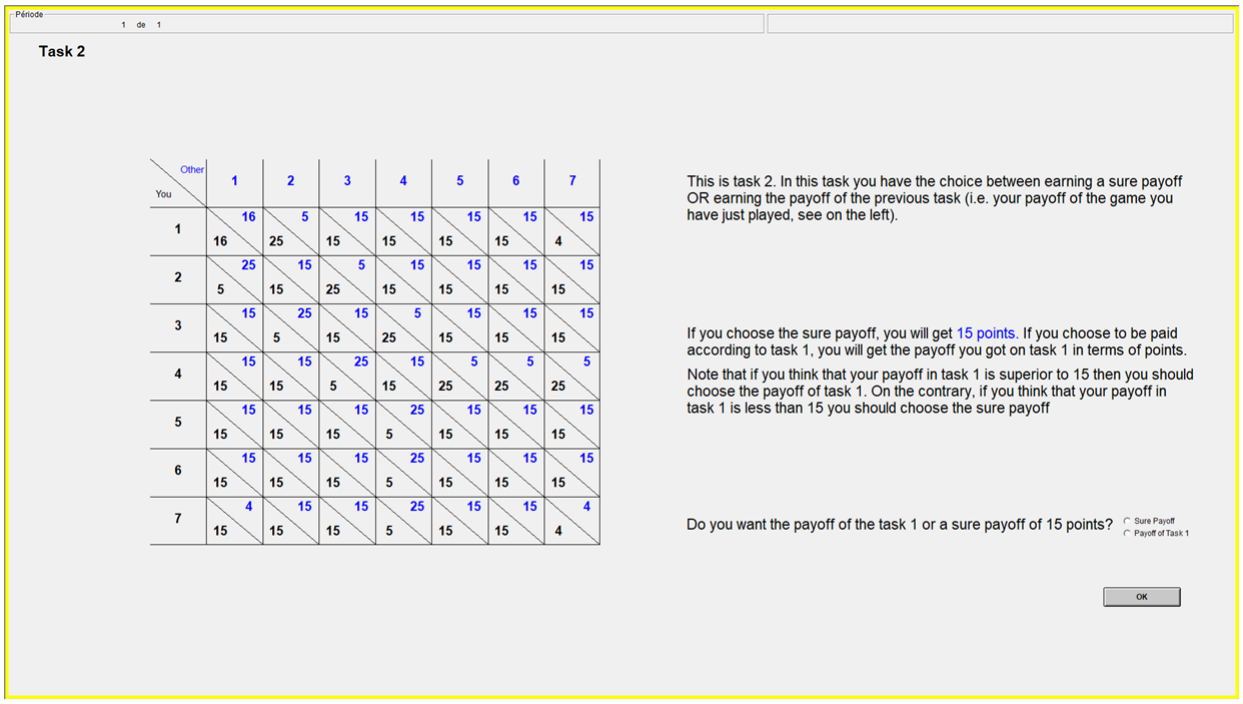


Figure 1.9: Task 2 (Undercutting Game: Stage 2)

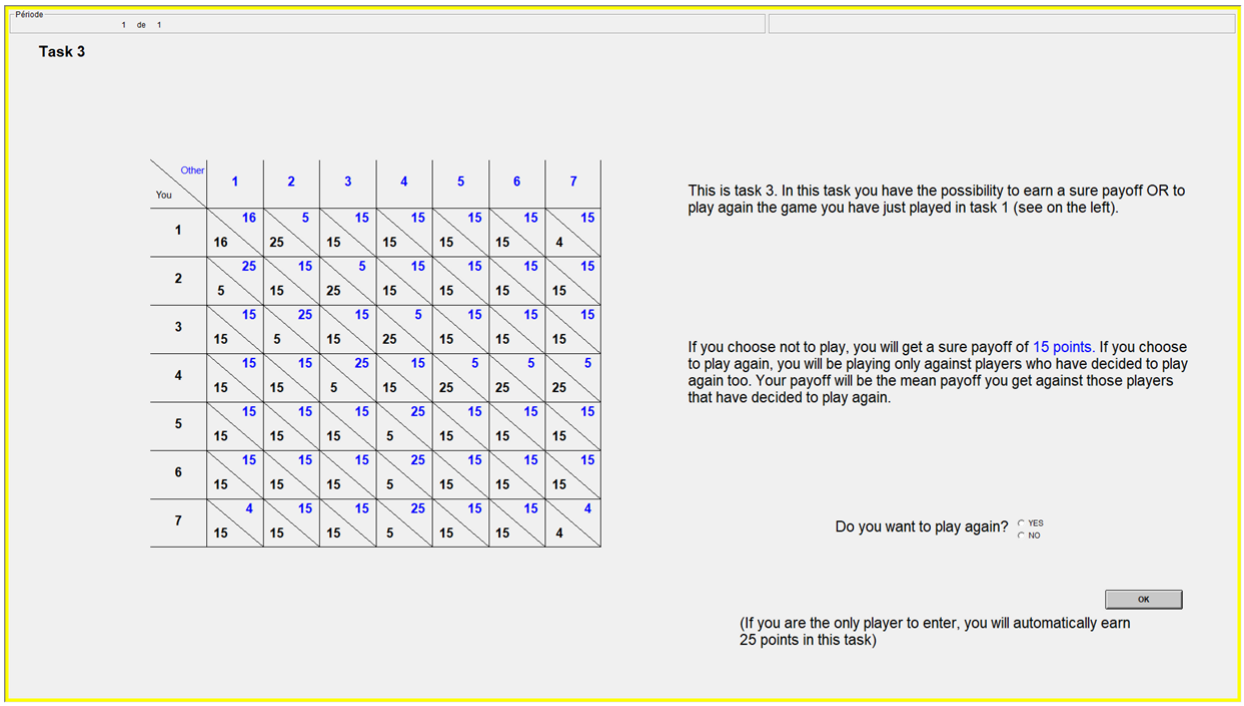


Figure 1.10: Task 3 (Undercutting Game: Stage 2)

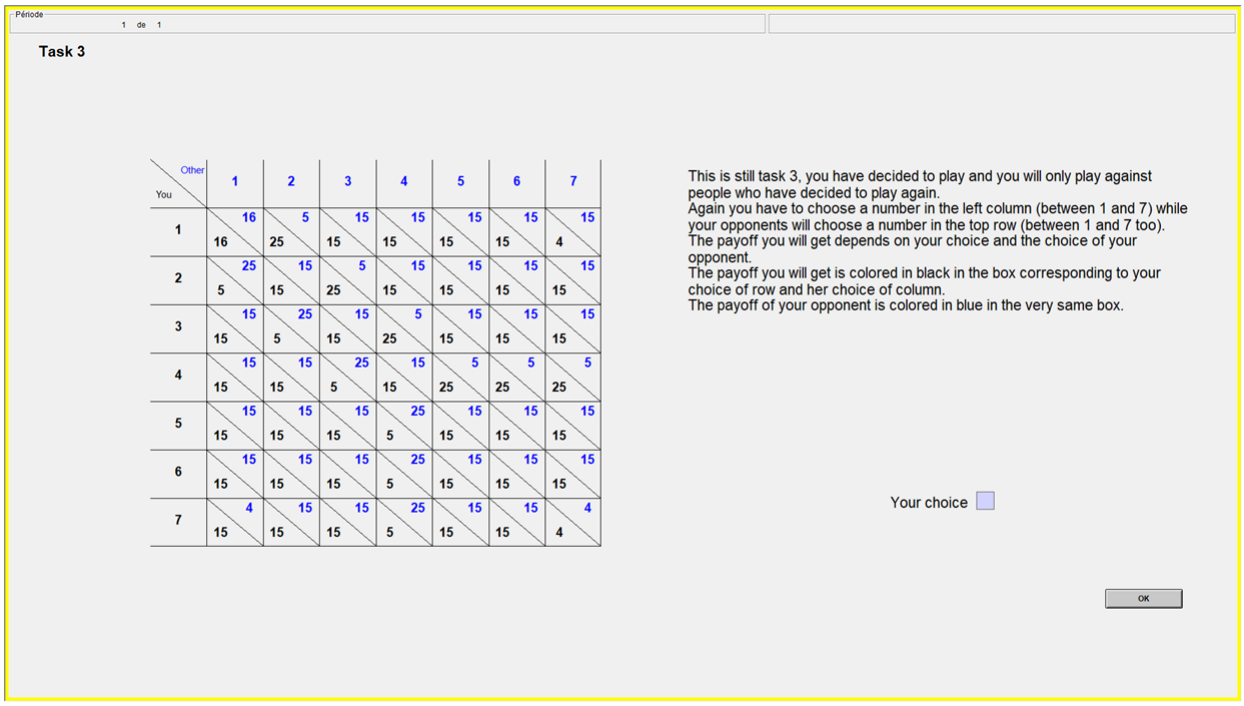


Figure 1.11: Task 3 (Undercutting Game: Stage 2)

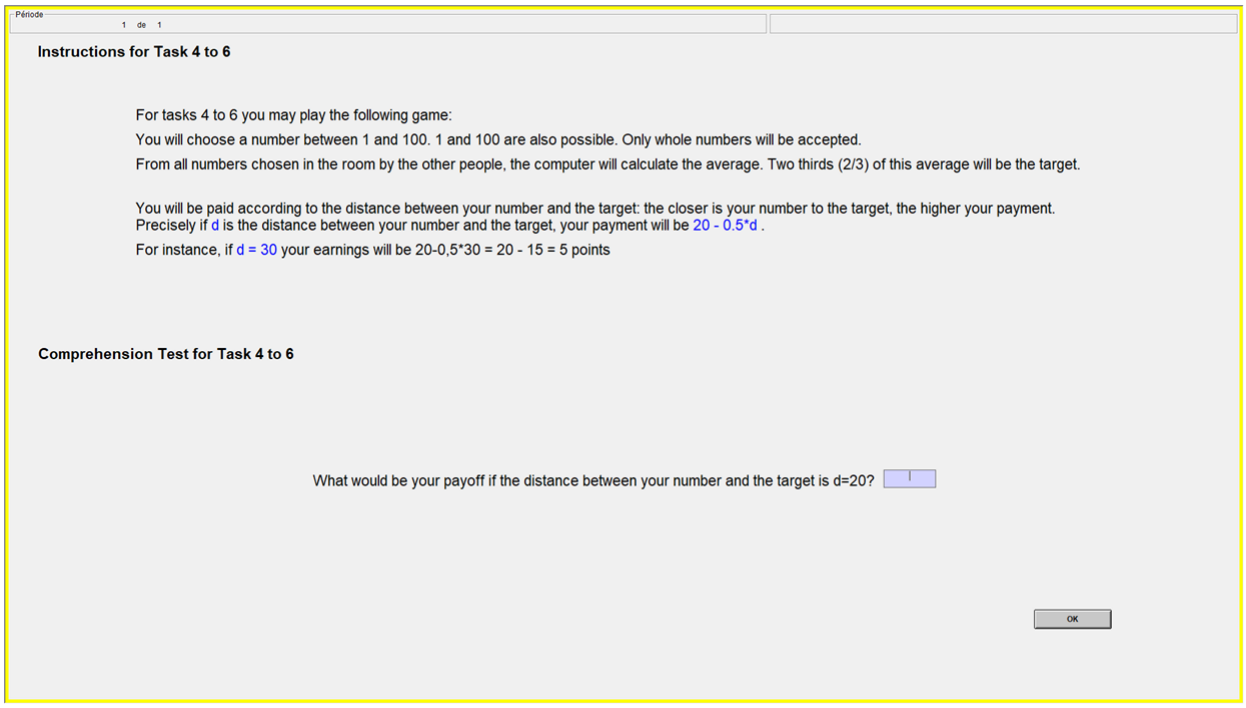


Figure 1.12: Instructions for Tasks 4-6

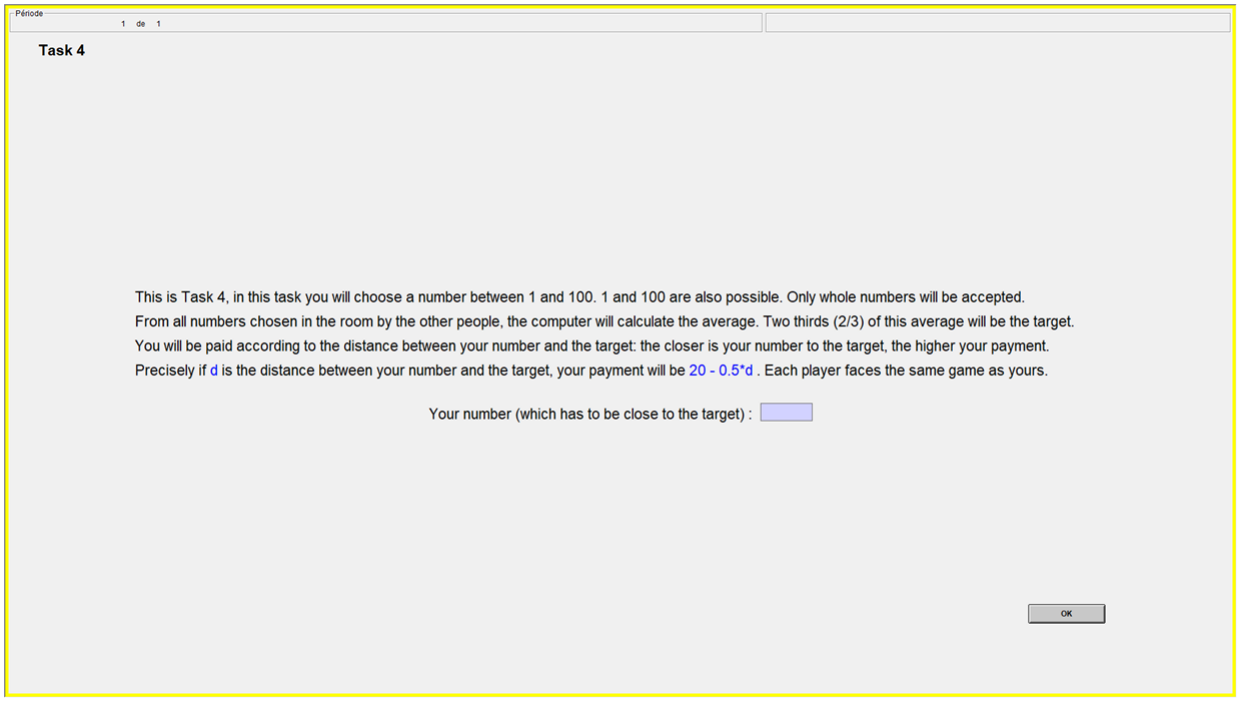


Figure 1.13: Task 4 (Beauty-Contest Game: Stage 1)

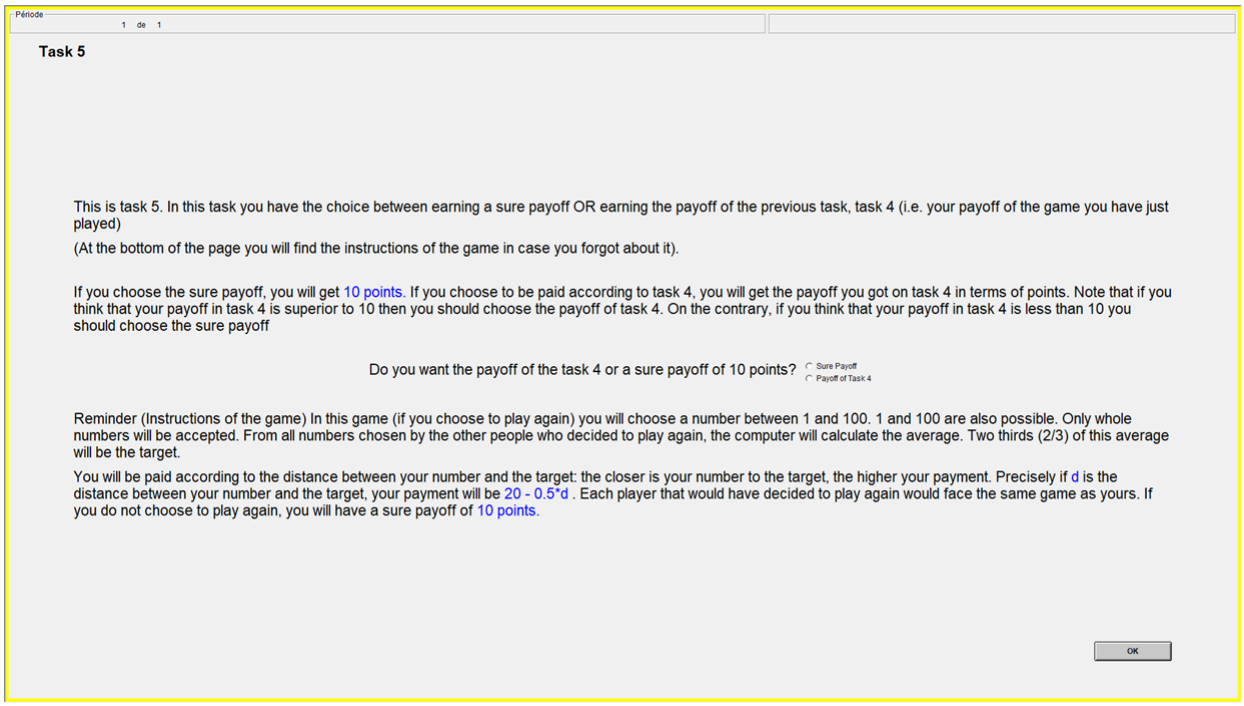


Figure 1.14: Task 5 (Beauty-Contest Game: Stage 2)

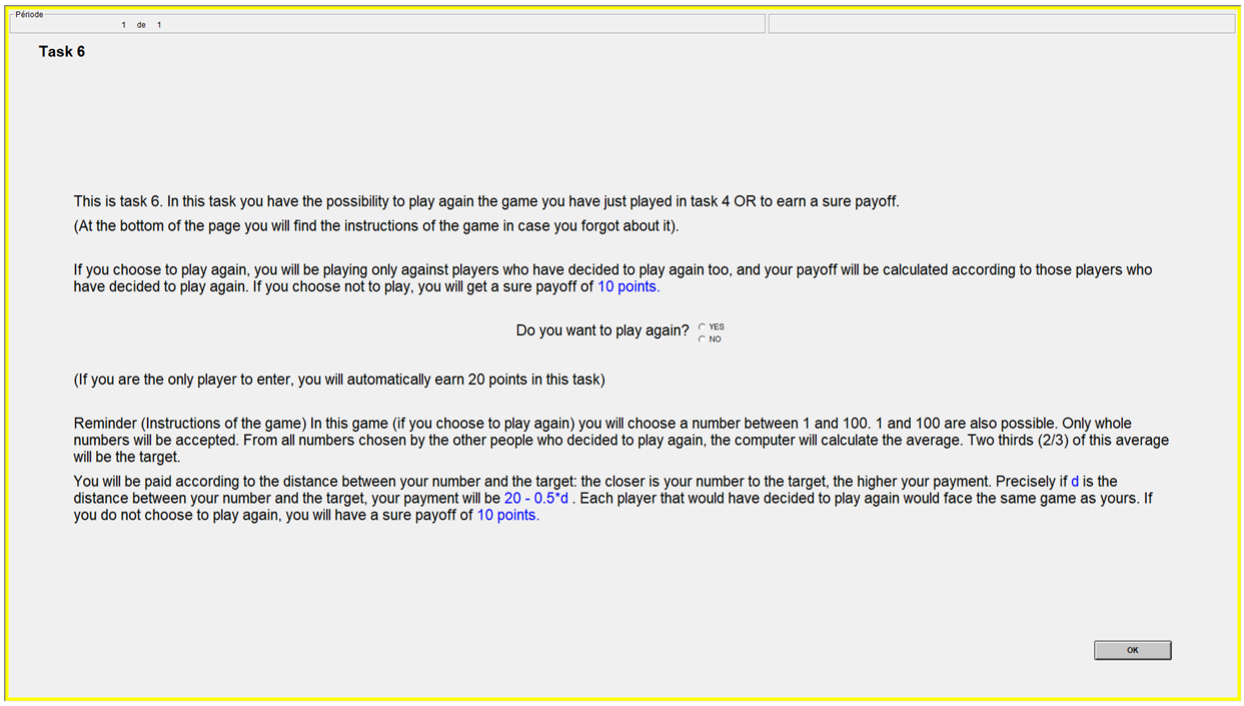


Figure 1.15: Task 6 (Beauty-Contest Game: Stage 3)

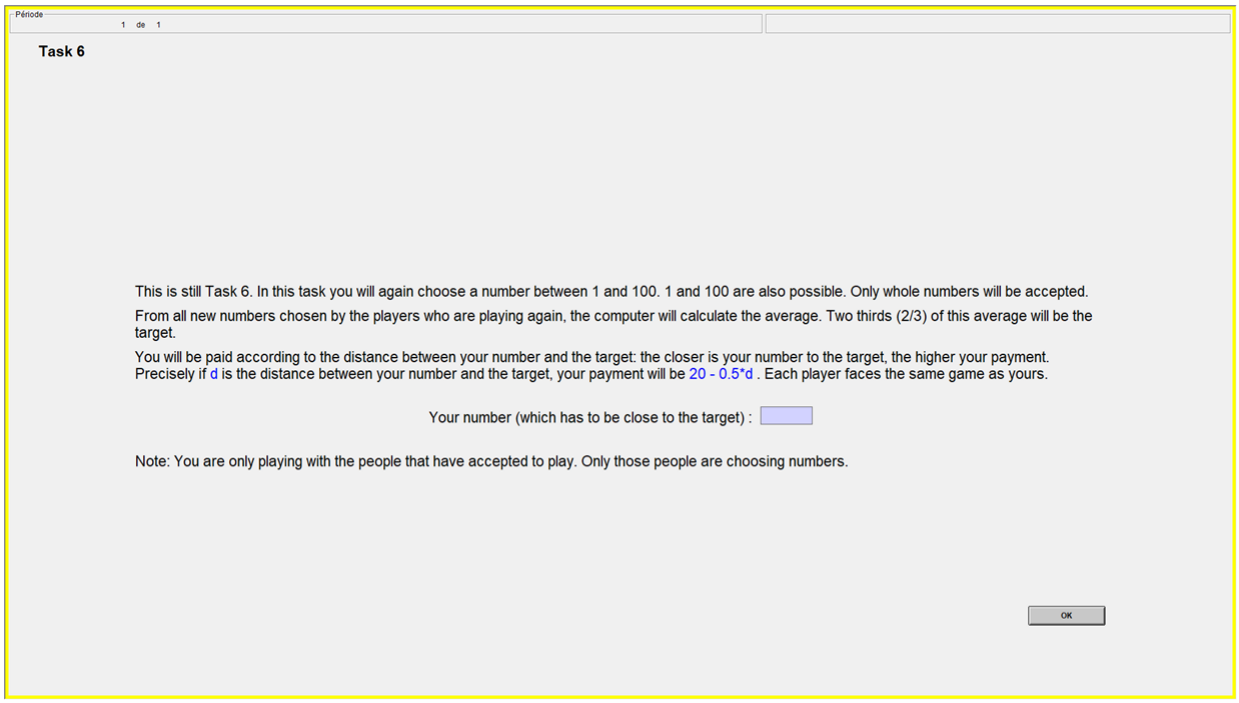


Figure 1.16: Task 6 (Beauty-Contest Game: Stage 3)

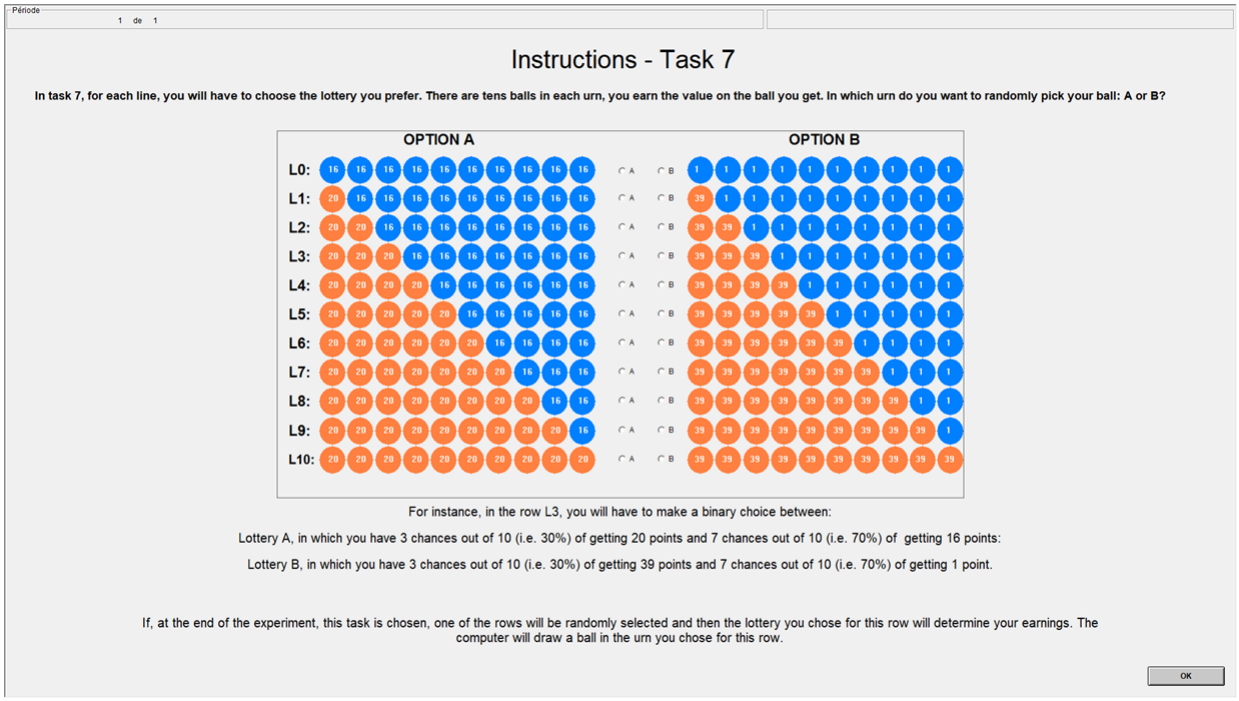


Figure 1.17: Task 7: Holt-Laury

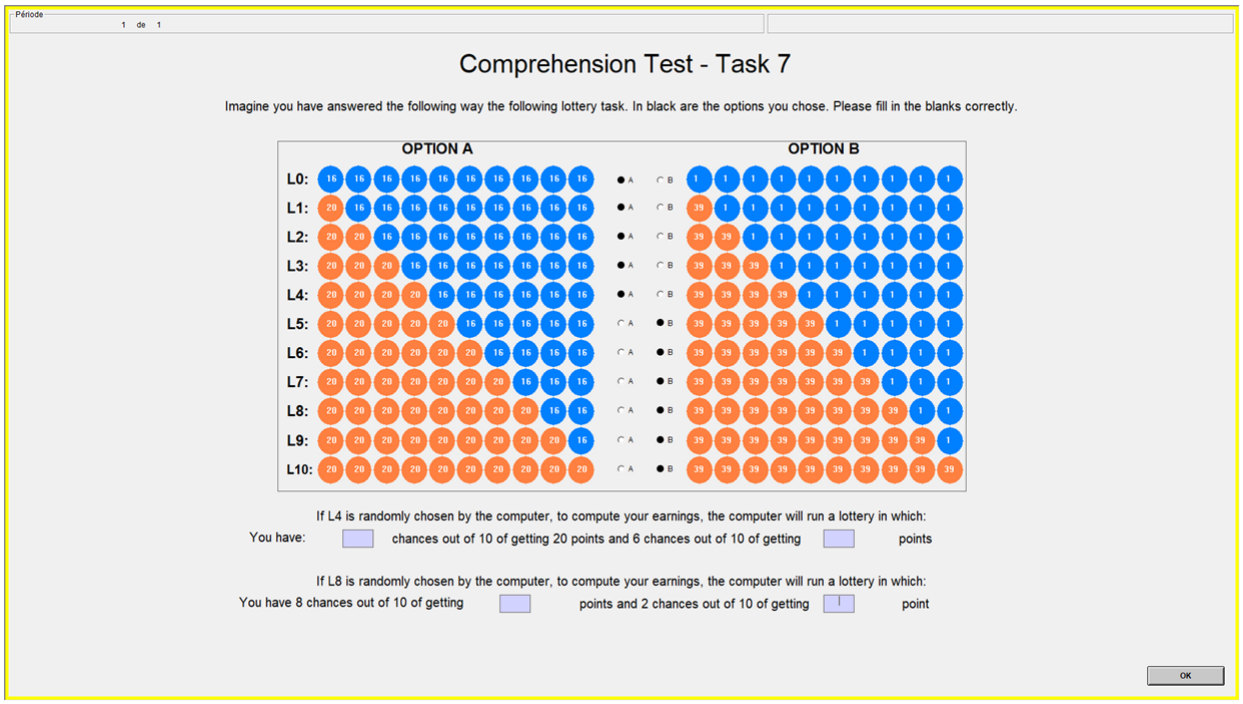


Figure 1.18: Task 7: Holt-Laury

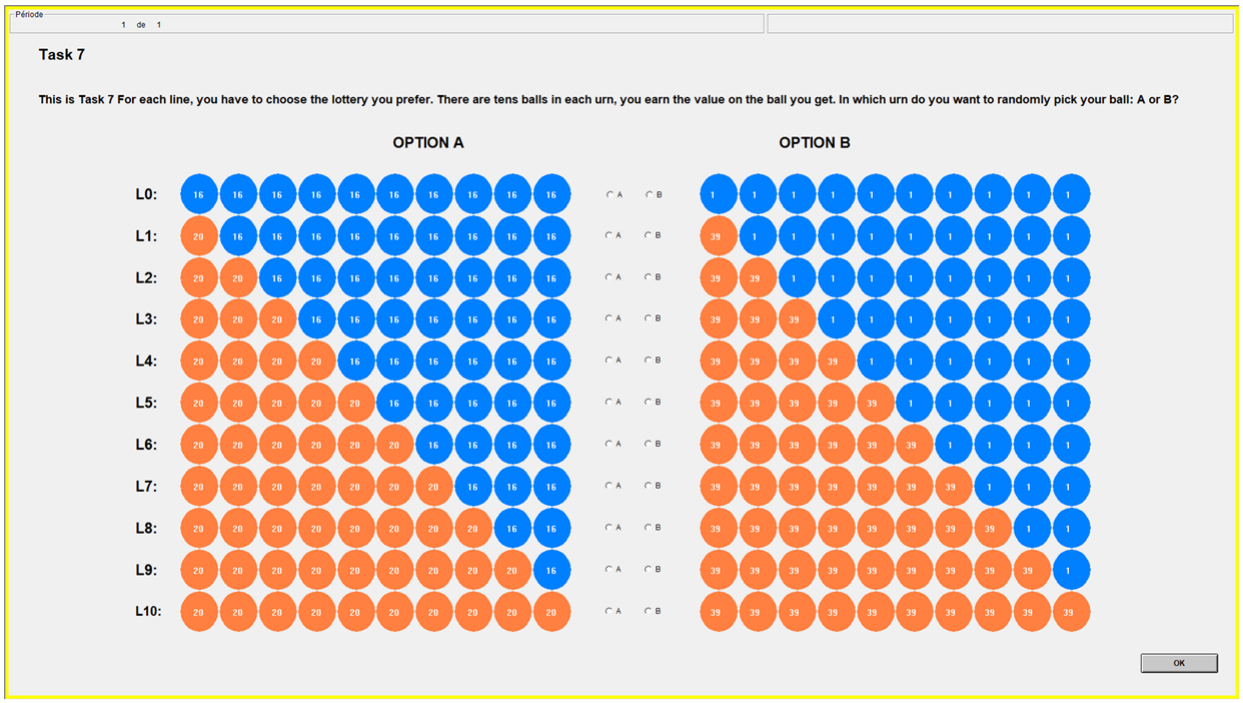


Figure 1.19: Task 7: Holt-Laury

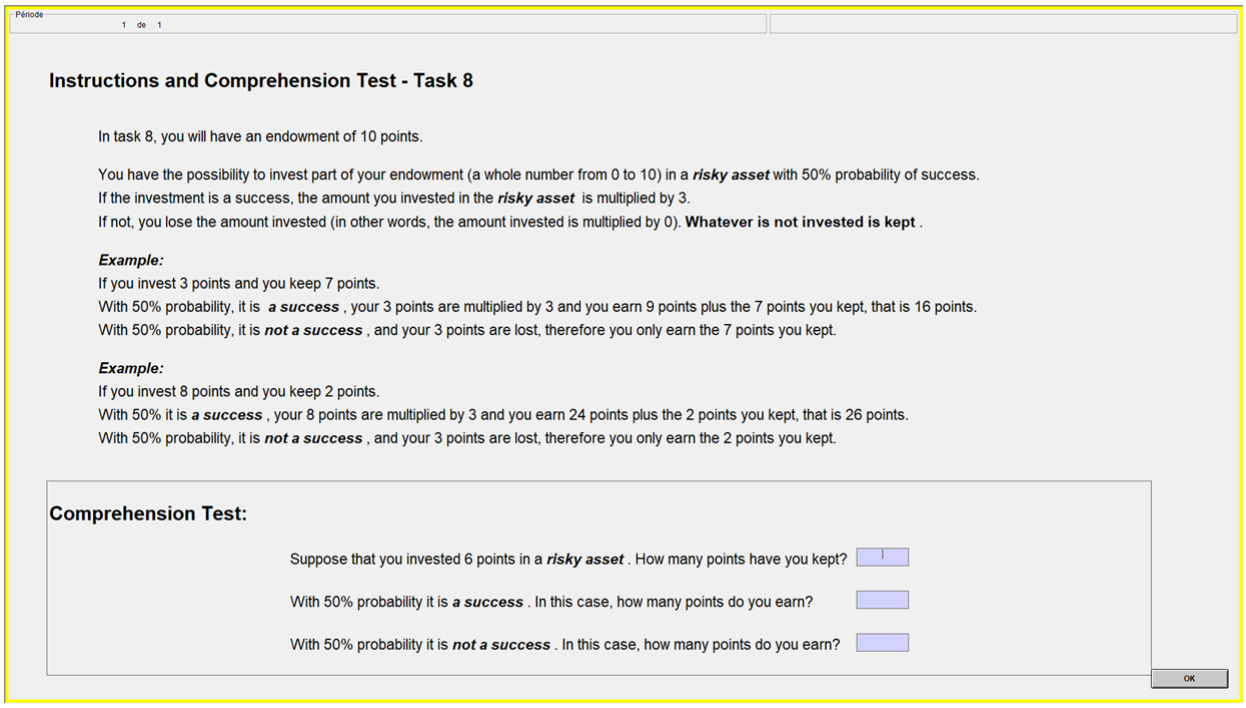


Figure 1.20: Task 8: Gneezy-Potters

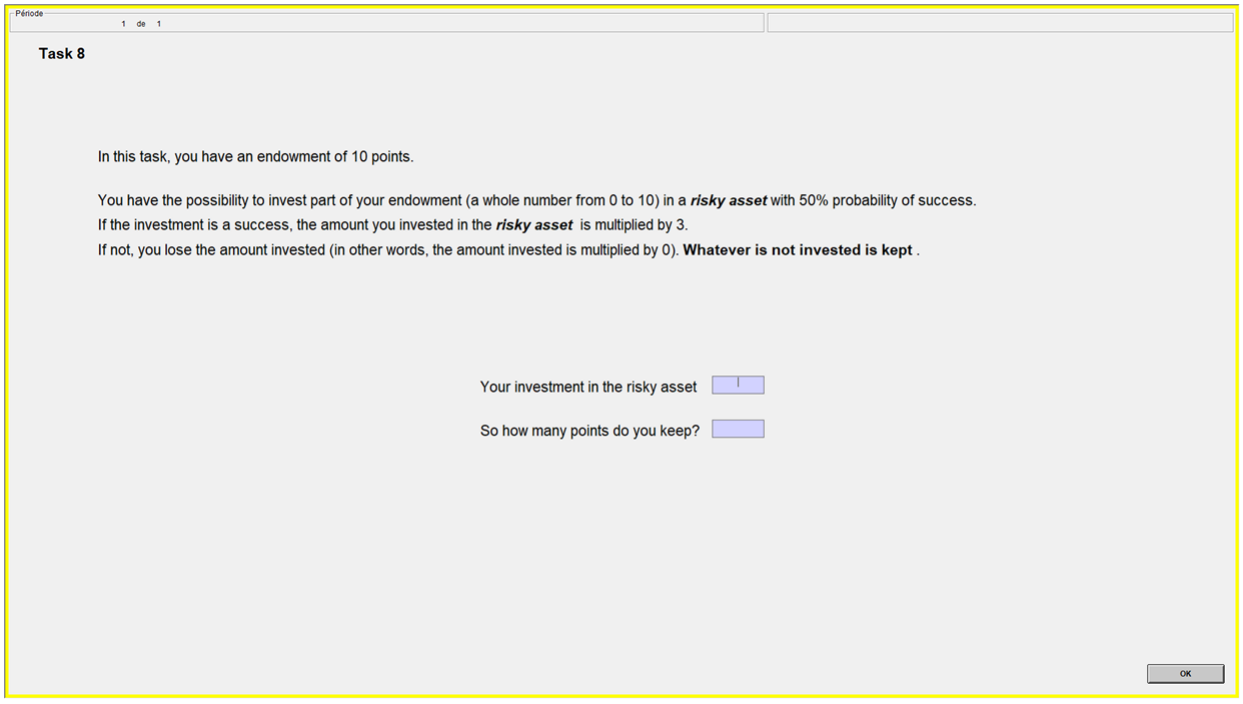


Figure 1.21: Task 8: Gneezy-Potters

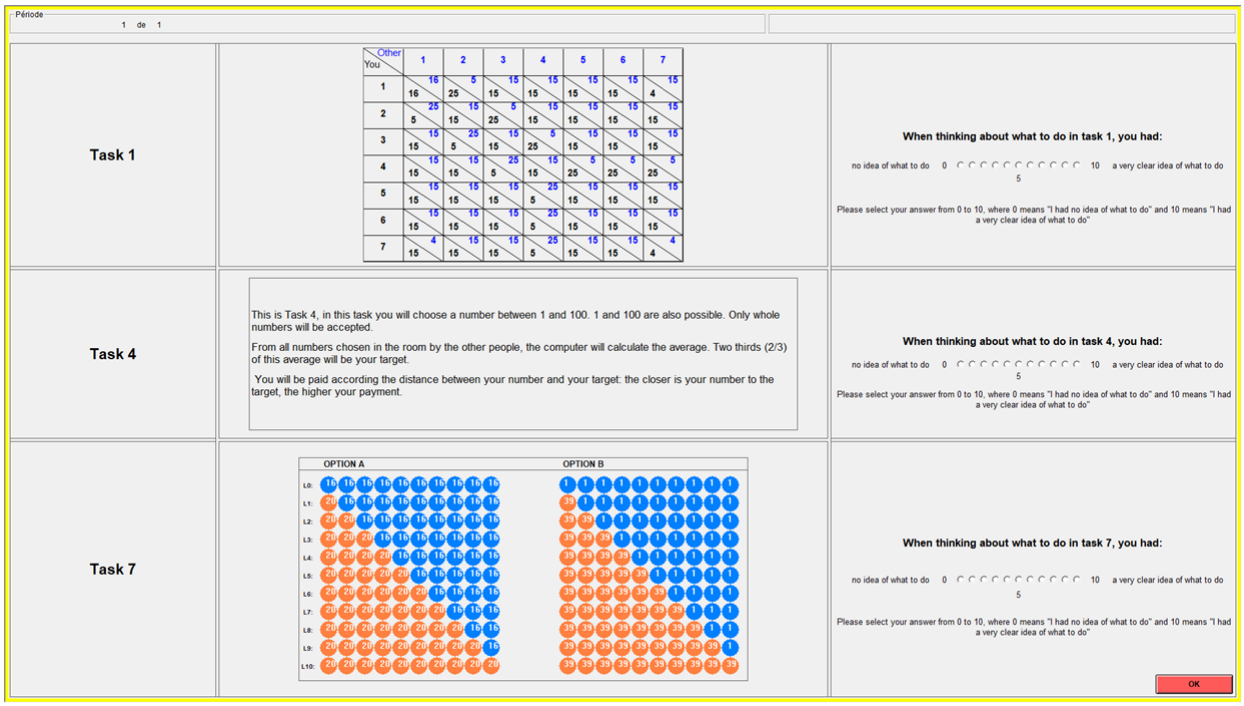


Figure 1.22: Task 9: Confidence

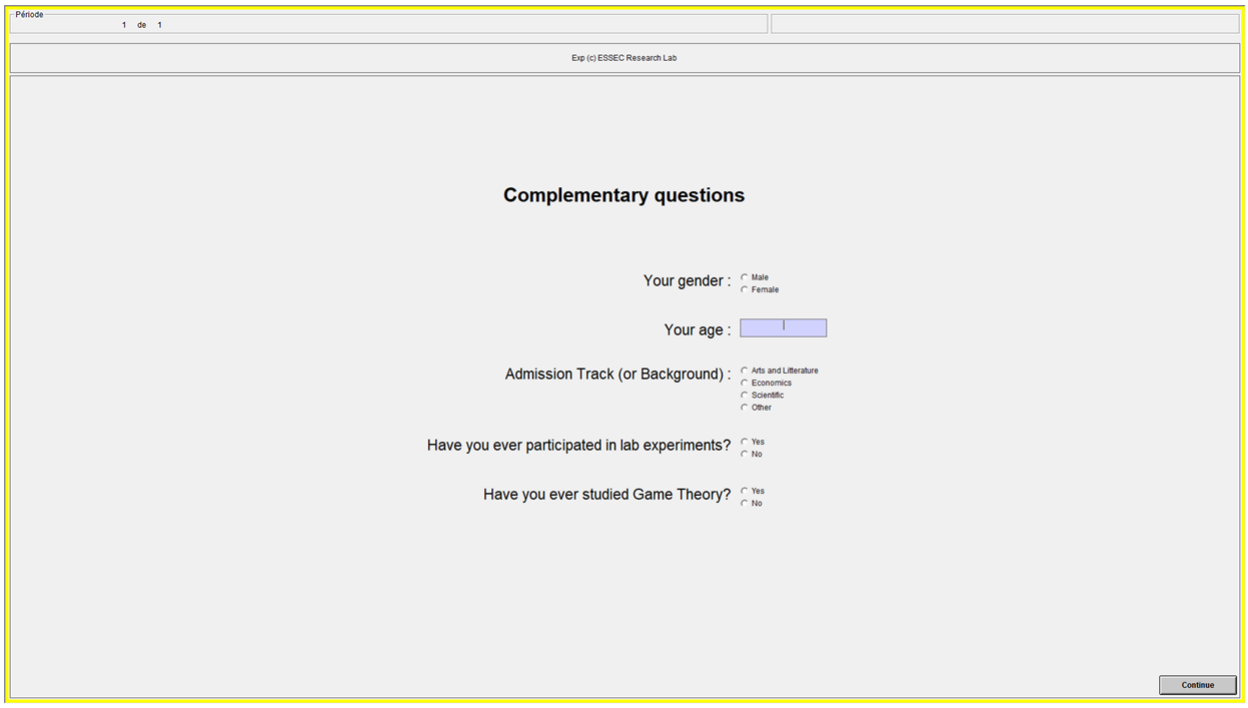


Figure 1.23: Demographics

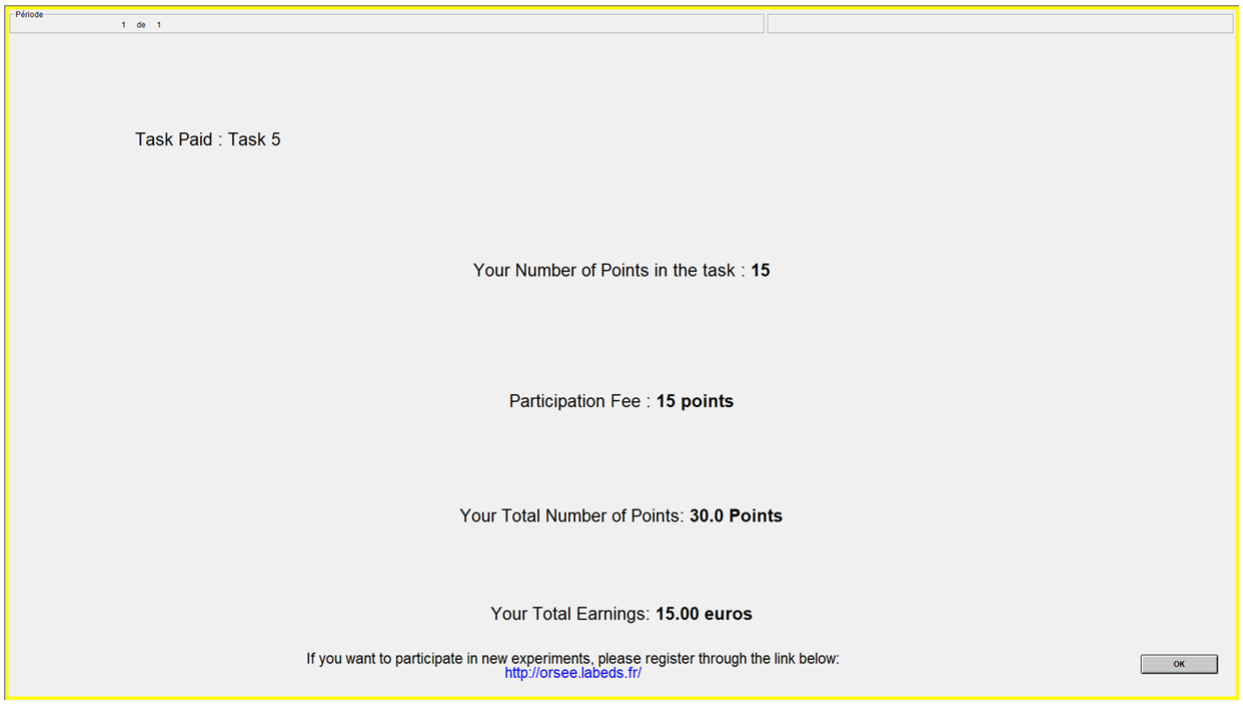


Figure 1.24: Payoffs

Chapter 2

Better Beliefs in Normal Form Games

Note: This chapter is co-authored with Guillaume Hollard.[1]

Abstract

Deviations away from equilibrium prediction in one-shot games are commonly observed in experimental games. Models were designed to capture the behavior of bounded-rational players in games. While the Quantal Response Equilibrium model allows players to *better respond* to correct beliefs, in the model we introduce, strategic players best-respond to *better beliefs*, beliefs that are not perfect. Unlike Level-k models, strategic players hold plausible beliefs that are (almost) correct on average. Specifically, our model assumes the existence of not forming beliefs (NFB) players that are either random or naive and the existence of strategic players. Strategic players best respond to their beliefs regarding the relative proportions of strategic and NFB players. Strategic players hold heterogeneous beliefs and fail to anticipate that they disagree on the relative proportion of NFB players. We show that our model has interesting properties and can be brought to the data by adding no more than one free parameter.

Keywords: Behavioral Game Theory; Model; non-cooperative games; beliefs.

[1]For this chapter I specifically would like to thank Gwen-Jiro Clochard, Heloise Cloléry, Philippe Choné, Jules Depersin, Laure Goursat, Nobuki Hanaki, Yukio Koryama, Dorothea Kubler, Rida Laraki, Philippos Louis, Vianney Perchet, Angela Sutan, and Georg Weizsäcker. I am also grateful to participants at he CWEE workshop 2019 at CREST

1. Introduction

Deviations away from Nash equilibrium predictions in one-shot games are routinely observed in experimental data (Goeree and Holt, 2001; Camerer, 2003). Although no clear consensus has emerged on the best alternative to the Nash solution concept, a lot has been learned on how to explain and predict observed strategies. We here propose a model that accounts for robust empirical findings while trying to avoid the most unrealistic features of existing models. A key feature of our model is to assume that players are heterogeneous along two dimensions. First, we acknowledge that players differ regarding their strategic ability. Therefore, we distinguish two types of players: *Not-forming beliefs* (NFB) players and *Strategic* players. NFB players can be either purely random or naively best-responding as if there were only random players[1]. NFB players form thus a mix of level-0 and level-1 players as defined in level-k models. The relative proportion of level-0 and level-1 players within the set of NFB players is likely to vary with characteristics of the games or subject pools. Strategic players form beliefs regarding the proportion $1-b$ of NFB players and the proportion b of strategic players and best respond to their beliefs. Second, we also acknowledge that strategic players may differ regarding their beliefs about the fraction of NFB players, i.e., they hold different b. Assuming heterogeneity in players' beliefs poses a difficult problem because players need to form beliefs about the distribution of b. We avoid this problem by introducing either players who do not form beliefs at all (the NFB players) or strategic players who consider that all other players have the exact same beliefs as theirs regarding the fraction of NFB players. In other words, strategic players wrongly believe that all strategic players hold the same b. However, the beliefs cannot be *too wrong* as we assume that values of b are correct on average. Strategic players realize there are non-strategic players and, on average, can gauge the proportion of such players.

Compared to the Quantal Response Equilibrium model that assumes that players *better respond* (i.e., tremble in the best-response process) to correct beliefs, our strategic players best respond to *better beliefs*, beliefs that are not perfect but correct on average. We thus name our model the *better beliefs model*.

Adding one more model to an already long list is a valid contribution if and only if some improvements are made regarding (at least some of) the following criteria: more realistic micro-foundations, greater ability to fit existing datasets, possibility to describe very different games by varying only one parameter (to avoid overfitting), ability to interpret variations of the free parameter. The better beliefs model meets some of the cited criteria. First, the model accounts for empirical facts that we consider essential, namely the existence of non-strategic players and the heterogeneity in beliefs. We provide a

[1]When payoffs are explicit, players behaving naively do not need to form beliefs to choose their strategy; they just maximize their expected payoff considering all the opponent's strategies as equally probable states of the world. That is why we include naive players in the NFB type of players.

one-parameter version of the model that can be brought to the data without further assumption. The free parameter depicts how strategic a subject pool facing a game is. We find that our model performs well at fitting existing data using data from two past studies (Georganas et al., 2015; Goeree et al., 2018). These datasets are composed of different games to provide a robustness check.

We first document some robust empirical regularities and compare the proposed model to existing ones in Section 2. We provide a simple example in Section 3 to offer insights into the intuition of our model. Section 4 introduces the theoretical properties of the model and Section 5 offers a parametric version of the model. Section 6 produces empirical estimations while Section 7 concludes.

2. Review

We first describe two empirical regularities that we consider as fundamental and thus encompassed in our model. We then review existing models in order to highlight our contribution.

2.1. Some facts

Existence of confused players. The existence of confused players is a matter of debate. For instance, in level-k models, level-0 players may only form a type of player that only exists in the mind of actual players. However, many experimental datasets show that even very poor strategies (e.g., dominated strategies) are sometimes played. It suggests that even meaningful strategies may have been played for reasons that have little to do with best-responding to reasonable beliefs (e.g., choosing one's lucky number). At the extreme, a player simply randomizing across possible strategies, using a uniform distribution, will often appear as playing sounded strategies. Consequently, several papers aim at tiring apart confused players and players who are using deliberate strategies. We identified six papers that differ in their methodology used to identify confused subjects: Costa-Gomes and Weizsäcker (2008); Agranov et al. (2012); Burchardi and Penczynski (2014); De Sousa et al. (2013); Agranov et al. (2015); Fragiadakis et al. (2016). All methods elicit a surprisingly high fraction of confused subjects in one-shot games, say between 30% and 50%. Our point here is not to discuss the fraction of confused players, nor their origin (e.g., poor instructions, low stakes, cognitive limitations). We believe that the existence of a non-negligible fraction of confused players is an empirical regularity across many games and many subject pools. Therefore, as explained below, our model will explicitly assume that some players are confused. Since we are agnostic regarding the proportion of such players, we allow our model to include all possible situations, from no confused players to all players being confused.

Heterogeneity in beliefs. The Nash equilibrium concept states that all players have correct beliefs. In Nash theory, different players who play the same role in a game should all have the same beliefs[2]; they should thus be homogeneous in beliefs. Since at least the works of Nagel et al. (1993) and Stahl and Wilson (1995), theoretical models have allowed players to hold heterogeneous beliefs. Heterogeneity in beliefs is supported by empirical evidence that elicited beliefs differ across players (see for instance Costa-Gomes and Weizsäcker, 2008; Fragiadakis et al., 2019). Heterogeneity in beliefs can be accounted for by introducing types that have different beliefs. For instance, in level-k models, level-2 players have different beliefs than level-1 players. Our model allows a different sort of heterogeneity since players with the same *type* (in other words, the same way of reasoning in games) can hold heterogeneous beliefs.

2.2. Behavioral game theory models

To situate more precisely our model in the behavioral game theory landscape, we provide a comparison with a dozen behavioral models regarding four criteria: heterogeneity in beliefs, existence of non-strategic/confused players in the model, ability to best-respond and inclusiveness ability (see Table 2.1). **Inclusiveness** refers to the existence of players who form beliefs about actions of others by mean of fixed-point reasoning "à la Nash". *Inclusiveness* is however more general than the standard Nash equilibrium in three ways: first *inclusiveness* may apply to a sub-group of players only. Second, inclusive players may also form beliefs about other players who do not include themselves in their reasoning. Last, *inclusiveness* is also used in the context in which strategies include a random element, as in the Quantal Response Equilibrium model. Level-k models (Stahl and Wilson, 1994, 1995; Nagel, 1995; Camerer et al., 2004; Alaoui and Penta, 2016) do not consider *inclusiveness*. Indeed, players are classified in different levels and only have the ability to best-respond to lower-level players[3].

The **existence of non-strategic players** criterion applies to models in which some players are assumed to not form beliefs at all. The **perfect best-response** criterion corresponds to the assumption that players are choosing their strategies within the set of best-response given their beliefs. Last, **heterogeneity in beliefs** refers to models which assume the possibility for players to hold heterogeneous beliefs.

[2]At least when the game possesses a unique Nash Equilibrium

[3]More precisely, in Alaoui and Penta (2016), players' depths of reasoning are endogenous and result from a cost-benefit analysis but players still best-respond to lower-level strategies

Table 2.1: Comparison of Behavioral models

	Heterogeneity of Beliefs	Existence of Non-strategic	Perfect BR	Inclusiveness
NI	×	×	×	×
QRE	×	×	×	✓
Nash	×	×	✓	✓
NLK	×	✓	✓	✓
Hetero QRE	✓	×	×	✓
ALE/ANNE	✓	×	×	✓
RBE/NBE	✓	×	×	✓
M-equilibrium	✓	×	×	✓
Hetero NI	✓	✓	×	×
Level-K (SW)	✓	✓	×	×
Level-K (N)	✓	✓	✓	×
CH	✓	✓	✓	×
ICH	✓	✓	✓	✓
Better Beliefs	✓	✓	✓	✓

BR refers as Best-Response, NI: Noisy Introspection (Goeree and Holt, 2004), QRE: Quantal Response Equilibrium (McKelvey and Palfrey, 1995), Nash: Nash equilibrium (Nash et al., 1950), NLK: NLK model (Levin and Zhang, 2019), Hetero NI: Heterogeneous Noisy Introspection (Goeree et al., 2018), Hetero QRE: Heterogeneous Quantal Response Equilibrium (Rogers et al., 2009) ALE/ANNE Asymmetric Logit Equilibrium/Asymmetric Noisy Nash Equilibrium (Weizsäcker, 2003), RBE: Random Belief Equilibrium (Friedman and Mezzetti, 2005) NBE: Noisy Belief Equilibrium (Friedman and Ward, 2019), M-equilibrium (Goeree and Louis, 2021) Level-k (SW) Level-k model (Stahl and Wilson, 1994, 1995) Level-K (N): Level-k model (Nagel, 1995), CH: Cognitive Hierarchy (Camerer et al., 2004), ICH: Inclusive Cognitive Hierarchy (Koriyama and Ozkes, 2021)

As can be seen from Table 2.1, there are substantial differences across models with respect to those four criteria. We precise below how the better beliefs model differ from the Inclusive Cognitive Hierarchy (ICH) model (Koriyama and Ozkes, 2021) and NLK model(Levin and Zhang, 2019) that are the closest to our model.

Both ICH and NLK models consider *inclusiveness* and the existence of confused subjects. In NLK model, strategic players best-respond to a mix of random and strategic plays as in our better beliefs model. But contrary to our approach, all strategic players in their model have the same belief regarding the proportion of non-strategic players while our strategic players hold heterogeneous beliefs regarding the proportion of NFB players. In ICH model, players with the same level of reasoning (as in Cognitive Hierarchy and Level-k models) all have the same beliefs regarding the proportion of other level-j players.

The heterogeneity in beliefs comes from the heterogeneity of levels. On the contrary, in our model, there is only one type of strategic players but they hold heterogeneous beliefs.

The idea of noise in beliefs is not new. Friedman and Mezzetti (2005) introduced Random Belief Equilibrium but rather to refine Nash equilibrium concept than to account for behaviors in the lab. More recently, Friedman (2018) and Friedman and Ward (2019) introduced and tested a class of models in which players are best-responding to noisy beliefs. Our approach is different since we assume the existence of two types of players and therefore the heterogeneity in beliefs is micro-founded by the approximate estimation (by a strategic player) of the probability of facing a strategic or a not-forming beliefs player.

3. A simple example: the p-beauty contest game

The way we model the reasoning of strategic players can easily be explained with the example of the p-beauty contest game. A p-beauty contest game consists in guessing the number closest to p times the average of all players' numbers. Imagine a subject, Anna, entering an experimental lab and being asked to play the 2/3-beauty contest game against students. Anna observes that some students have never learned game theory and do not seem at ease when choosing their strategies, while others seem to have some experience. Anna estimates the proportion of students in each of the two categories above. She decides to answer as if a proportion $1 - b$ of students were choosing randomly or naively their number and a proportion b of students were best-responding as she is. Thus, the number chosen by Anna is thus neither too high since she thinks there is a significant proportion of strategic players nor too low owing to her relative estimation of the share of confused subjects. The only problem is that Anna does not know precisely how many strategic players there are and ignores that they may not have the same beliefs regarding the proportion of strategic players. Still, Anna's beliefs are not that bad, and she is not entirely off the mark. Anna is the representation of a strategic player in our model.

In the 2/3 beauty contest game, confused players will either play randomly (and thus pick a random number between 0 and 100) or naively (play 33 as if all numbers were chosen randomly). Strategic players hold heterogeneous b and therefore could play numbers between 0 and 33 depending on their beliefs. In Figure 2.1 we materialize this graphically with the plays predicted by a parametric version of our model that we will describe in more depth below[4]

[4]We provide a first glance of the predictions of our model by choosing the parameter that matches the mean play predicted by our model with the empirical mean from the Financial Times experiment. We do the same below with Logit-QRE and Poisson Level-K models.

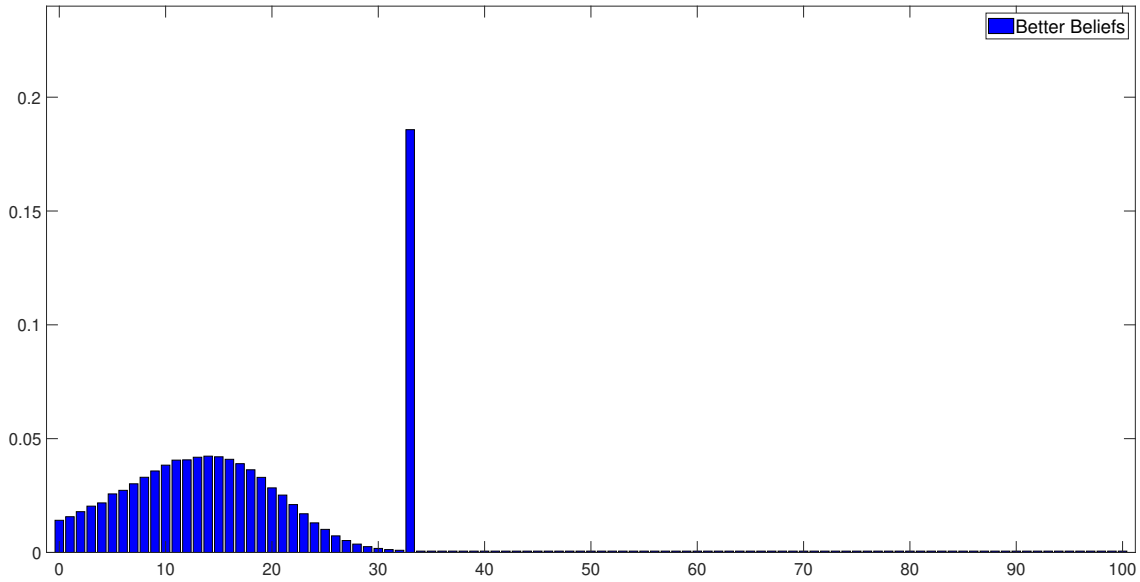


Figure 2.1: Predicted strategies by a parametric version of the better beliefs model in the 2/3-beauty contest game

Below can be found graphically what would be the predicted strategies by Logit-QRE model and Level-k model with Poisson distribution of levels in Figure 2.2 and Figure 2.3. We observe that those three models are different in their predictions. The empirical distribution from the Financial Times newspaper experiment can be found in Figure 2.4.

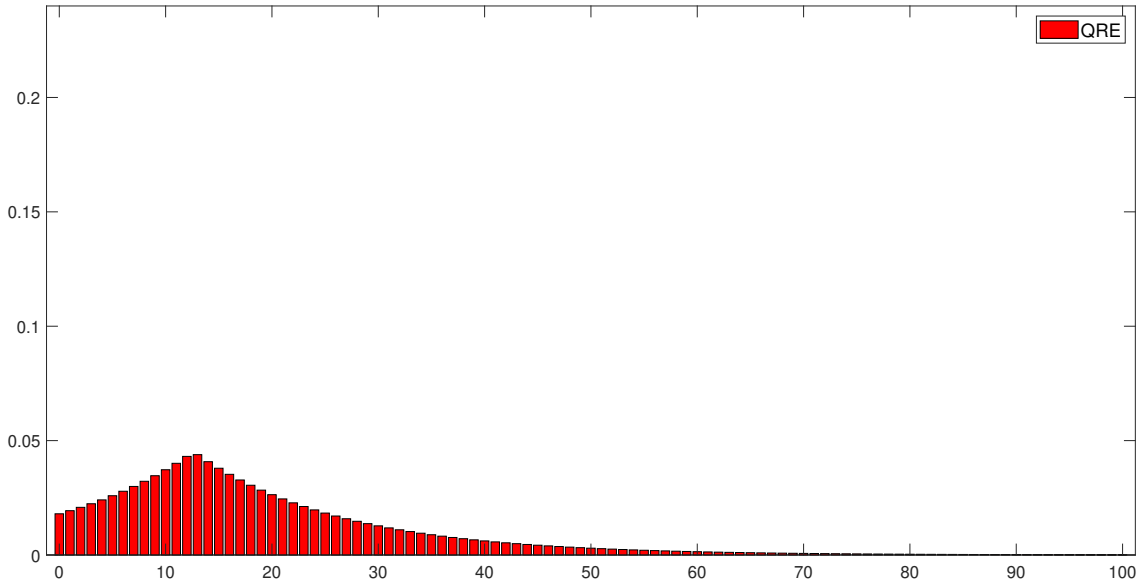


Figure 2.2: Predicted strategies by the Logit-QRE model in the 2/3-beauty contest game

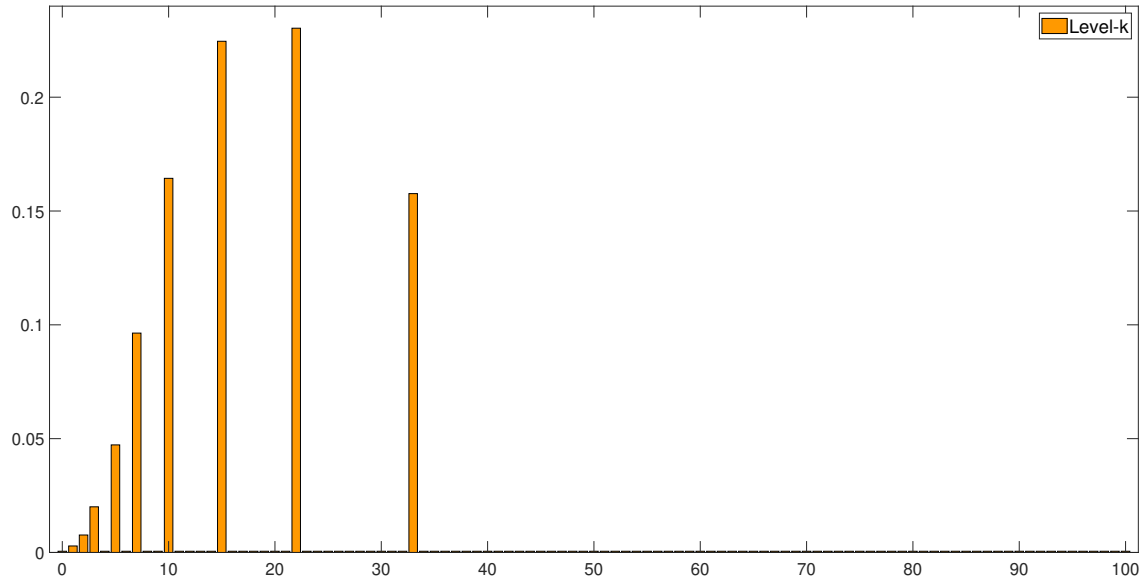


Figure 2.3: Predicted strategies by the Poisson Level-k model in the 2/3-beauty contest game

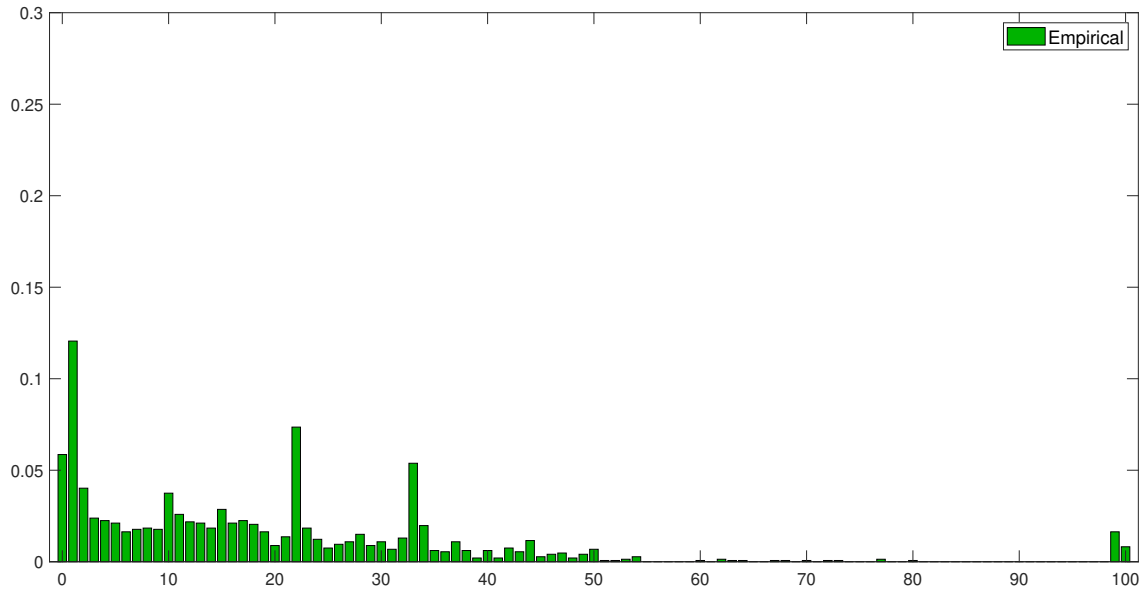


Figure 2.4: Empirical strategies in the 2/3-beauty contest game (Financial Times Experiment)

We see that differences in each model's underlying assumptions lead to different predictions (note that each presented model has one free parameter adjusted to fit the empirical mean). The better beliefs model share with the level-k models the possibility to get a spike corresponding to level-1. It shares with the QRE the possibility to have a continuum of strategies, with a significant proportion of players playing 0 or 1, which belong to the set of Nash equilibria.

4. The better beliefs model

4.1. General Framework

We consider finite normal form games. A finite normal form game Γ is a triplet $\{N, S, u\}$:

- N is the set of players, $N = \{1, .., n\}$
- S is the strategy space, $S = \prod_{i=1}^{n} S^i$ with $S^i = \{s_1^i, ..., s_{k_i}^i\}$ (player i can choose among k_i pure strategies)
- u is the payoff functions vector, $u = (u^1, ..., u^n)$ with $\forall i \in N, u^i : S \to \mathbb{R}$

Let X be the extended strategy space that contained mixed strategies.

$$X = \prod_{i=1}^{n} X^i \text{ with } X^i = \{(x_1^i, ..., x_{k_i}^i) \in \mathbb{R}_+^{k_i}, \sum_{j=1}^{k_i} x_j^i = 1\}$$

x_j^i refers to the probability of player i to play strategy s_j^i. Let s_j^i be the canonical vector in X_i corresponding to the strategy s_j^i so that S^i is identified as a subset of X^i (hence a subset of $\mathbb{R}^{k_i}$).

X is a convex subset of $\prod_{i=1}^{n} \mathbb{R}^{k_i}$. If x and y are two elements in X, $\forall \lambda \in [0, 1]$ we have with the usual operations in R^n, $z = \lambda x + (1 - \lambda)y \in X$ with $z_j^i = \lambda x_j^i + (1 - \lambda)y_j^i$.

We extend the payoff functions vector $u = (u^1, ..., u^n)$ to X

$$\forall i \in N,\ u^i : X \to \mathbb{R} \text{ with } u^i(x) = \sum_{(j_1,...,j_n) \in [\![1,k_1]\!] \times ... \times [\![1,k_n]\!]} x_{j_1}^1 ... x_{j_n}^n u^i(s_{j_1}^1, ..., s_{j_n}^n)$$

We define for $i \in [\![1, n]\!]$ the individual best-response set-valued function $BR^i : X^{-i} \to X^i$ such that

$$\forall x^{-i} \in X^{-i}, BR^i(x^{-i}) = \{x^i \in X^i \mid u^i(x^i, x^{-i}) = \max_{z^i \in X^i}, u^i(z^i, x^{-i})\}$$

and we also define for $i \in [\![1, n]\!]$ the individual pure best-response set-valued function $PBR^i : X^{-i} \to S^i$ such that

$$\forall x^{-i} \in X^{-i}, PBR^i(x^{-i}) = \{s^i \in S^i \mid u^i(s^i, x^{-i}) = \max_{z^i \in S^i}, u^i(z^i, x^{-i})\}$$

We then define the global best-response set-valued function $BR : X \to X$ such that

$$\forall x \in X, BR(x) = \{y \in X \mid \forall i \in [\![1, n]\!]\ y^i \in BR^i(x^{-i})\}$$

4.2. Two types of players

In the model, we assume that they are two types of players: not-forming beliefs (NFB) players and strategic players.

Not-forming beliefs (NFB) players

Depending on the framing of the game (matrix normal form game, instructions without clear relation between plays and payoffs) and on their level of confusion, NFB players tend to act randomly (we will denote them *rand* and assume that they play randomly) or naively (we will denote them *naive* and assume that they best-respond to random play).

$$\forall i \in N \ \sigma^i_{rand} = \frac{1}{k_i} \sum_{j=1}^{k_i} s^i_j$$

$$\forall i \in N \ \sigma^i_{naive} \in BR^i(\sigma^{-i}_{rand})$$

Precisely, to make *naive* players distribution of plays deterministic, we set:

$$\forall i \in N \ \sigma^i_{naive} = \frac{1}{\#(PBR^i(\sigma^{-i}_{rand}))} \sum_{s^i_j \in PBR^i(\sigma^{-i}_{rand}))} s^i_j$$

When payoffs are explicit, *naive* players do not form beliefs to choose their strategy they just maximize their expected payoff considering all the opponent's strategies as equally probable states of the world.

Strategic players (b-Nash players)

Strategic players adapt their behavior to their beliefs about the size of the NFB players' pool and hold heterogeneous beliefs. More precisely, a strategic (also denoted as b-Nash) player i assumes that a proportion of players $1-b_i$ is NFB while a proportion of players b_i is strategic. She believes that b_i is common knowledge across strategic players. Furthermore, b-Nash players also form beliefs regarding the strategies of those NFB players and thus about the fraction of *naive* and *rand* players inside the NFB pool of players. We will precise this below. For now, we denote $\sigma_N(b_i)$ the believed NFB aggregated strategy.

Strategic players thus play *à la Nash* considering that all strategic players hold the same belief. Extreme values regarding b are interesting to comment: if b is equal to 1, the player chooses a Nash strategy while if b is equal to 0, the player best responds to NFB players.

As in Nash equilibrium concept, the definition of the strategy played by a b-Nash player is a fixed-point solution[5]:

[5]This is the same fixed-point problem as in Levin and Zhang (2019)

$$\forall i\ \sigma^i(b) \in BR^i[\,(1-b)\,\sigma_N^{-i}(b)\ +\ b\,\sigma^{-i}(b)\,]$$
$$\sigma(b) \in BR[\,(1-b)\,\sigma_N(b)\ +\ b\,\sigma(b)\,] \tag{2.1}$$

Proposition 1: In every finite normal form game Γ, for every $b \in [0,1]$ and $\sigma_N(b) \in X$, there exists a solution to the fixed-point problem of a b-Nash player.

Proof is in Appendix A and relies on Kakutani Fixed-Point Theorem.

***Better beliefs*: distribution of b, the believed proportion of strategic players**

Strategic players have different beliefs regarding b, the proportion of strategic players. We denote β the *true* proportion of strategic players. We assume that strategic players have beliefs b close to β but not perfect, hence the notion of *better beliefs*. Precisely, we assume that for each strategic player i:

$$b_i = \beta + \epsilon_i \qquad \text{with } \epsilon_i \sim G$$

ϵ_i is the error made by a b-Nash player when estimating the size of the strategic pool of players (or the probability to face a strategic player). We will exhibit a parametric distribution of error G in next subsection to estimate the model.

5. A parametric version of the model

In this section, we provide a parametric version of the better beliefs model.

5.1. Fraction of b-Nash and not-forming beliefs players aggregated strategies

β is the fraction of b-Nash players and thus $1-\beta$ is the fraction of NFB players. We assume that among NFB players, the share of *naive* players is β, and that of *rand* is $1-\beta$. In other words, conditionally on not forming beliefs, the probability to be *naive* (respectively *rand*) is β (respectively $1-\beta$). The idea behind this parametrization is that the more confused players there are, the more random they behave.

Therefore, the NFB aggregated strategy is:

$$\sigma_N = \beta\sigma_{naive} + (1-\beta)\sigma_{rand}$$

5.2. Strategic players' beliefs regarding the NFB strategy

Since b is the noisy signal of β for strategic players, we assume that a b-Nash player will believe that inside the NFB pool of players:

$$\sigma_N(b) = b\sigma_{naive} + (1-b)\sigma_{rand}$$

Therefore we can be more precise regarding the fixed-point problem a b-Nash player solves:

$$\sigma(b) \in BR[\,(1-b)\,(b\sigma_{naive} + (1-b)\sigma_{rand}) \;+\; b\,\sigma(b)\,]$$

Definition (b-Nash strategy): $\sigma(b) \in X$ is a b-Nash strategy if it meets the condition above.

Remark 1: For all $b \in [0,1]$ there exist at least one b-Nash strategy. Indeed, it is a particular case of Proposition 1, with $\sigma_N(b) = b\sigma_{naive} + (1-b)\sigma_{rand}$.

Remark 2: β-Nash players (b-Nash players with $b = \beta$) have perfect beliefs regarding the proportion of strategic players and also regarding the composition of NFB players. They just ignore that strategic players are heterogeneous in terms of beliefs b.

5.3. Parametric distribution of b

We assume that for each strategic player i:

$$b_i = \beta + \epsilon_i \qquad \text{with } \epsilon_i \sim \underline{\mathcal{N}}(0, \nu^2)$$

where $\underline{\mathcal{N}}(0,\nu^2)$ is a truncated normal distribution so that the belief $b_i \in [0,1]$ (and so $\epsilon_i \in [-\beta, 1-\beta]$). The normal distribution is the most natural distribution for errors in beliefs. With this form, b-Nash players have a belief b close to β (the true proportion of strategic players).

Variance of the beliefs

To parameterize our model, we assume that the variance of the truncated normal distribution of errors is proportional to $\beta(1-\beta)$ ($\nu^2 = k\beta(1-\beta)$). The idea behind is that the more extreme - in term of proportion of NFB players - is a pool of players, the more precisely one can assess its composition.

If $k >> 1$, the distribution of beliefs is close to uniform. On the contrary if $k << 1$, the distribution of beliefs tends to a Dirac with all players having perfect beliefs. Therefore

we set $k = 1/4$, neither too low nor too high so that players have *better beliefs*[6].

Let F_β be the cumulative distribution of beliefs b. In Figure 2.5 you can find the distribution of beliefs for $\beta = 0.6$.

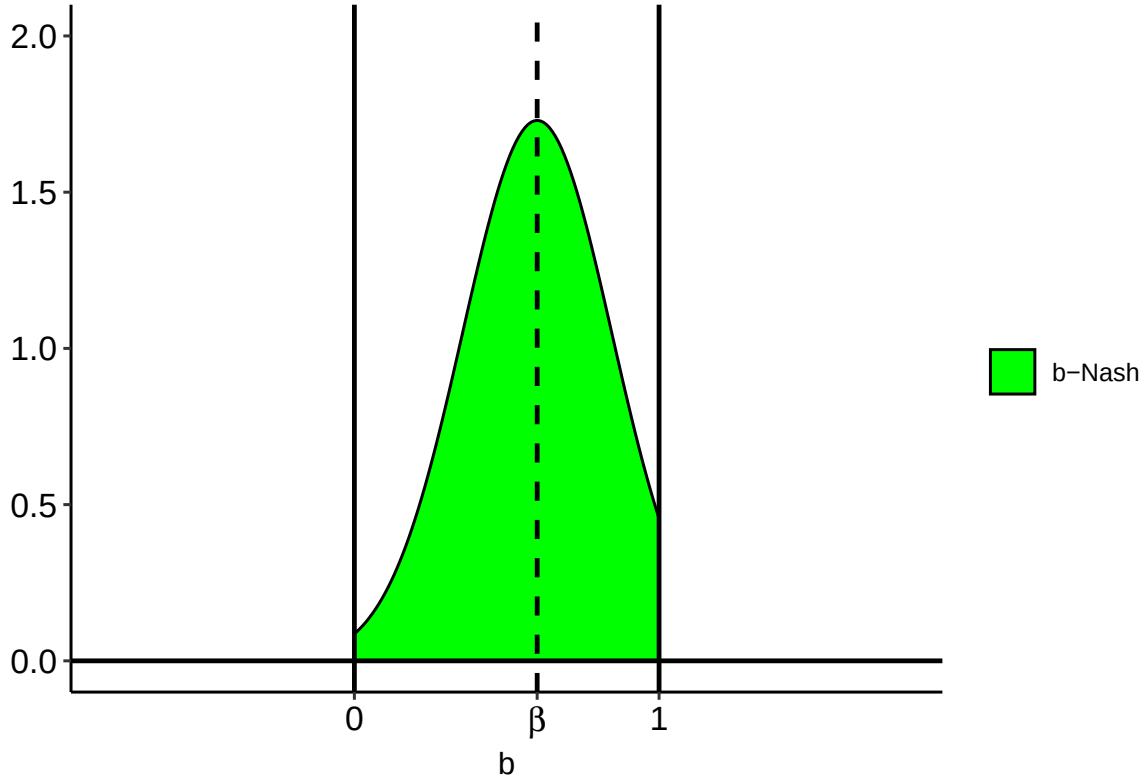


Figure 2.5: Distribution of beliefs for $\beta = 0.6$

5.4. Aggregated strategic players behavior

To compute the aggregated strategy played by b-Nash players, we need to be sure that we can integrate a selection of b-Nash strategies along the distribution of beliefs.

Proposition 2: For every normal form game, there exists a piecewise continuous function $\sigma_s : [0, 1] \to X$ such that $\sigma_s(b)$ is a b-Nash strategy.

Proof is in Appendix A and relies on semi-algebraic geometry, in particular on a proposition from Schanuel et al. (1991).

Given all the definitions notations and assumptions, a representative b-Nash strategy is:

$$S_s = \int_0^1 \sigma_s(b)\, dF_\beta(b)$$

[6]In Appendix D, we show that estimations do not change much with $k = 1/2$ and $k = 1/8$

with σ_s a piecewise continuous function as defined in Proposition 2 and where $\int$ extends the integration to X by means of coordinate-wise integration.

Remark 3: It can happen that more than one function σ_s can meet the conditions of our model. Therefore, as it can be the case in other models (Quantal Response Equilibrium models and Nash theory for instance), more than one distribution of strategies can meet the conditions of our model.

5.5. Repartition of types and subtypes

Now that we have defined all concepts and notions, note that β is the only parameter of the model. The distribution of aggregate strategy of the whole pool of subjects in function of β is:

$$\sigma(\beta) = \beta \int_0^1 \sigma_s(b)\, dF_\beta(b) + (1-\beta)[\beta\sigma_{naive} + (1-\beta)\sigma_{rand}]$$

β represents the proportion of strategic players inside a pool of subjects dealing with a game. In other words, it depicts how strategic a pool of subjects is when dealing with a specific game. The closer β is to 1, the more strategic and rational the pool of subjects. On the contrary, the closer to 0, the more confused the pool of subjects. We do not necessarily impose that β is the same across games for the same pool of subjects since some games may appear more strategically complex for some players. Indeed, some players can act strategically in a game but naively or randomly in another (see Hollard and Perez, 2020, for a discussion on instability of types across games).

In Figure 2.6 you can find the repartition of subtypes as a function of β.

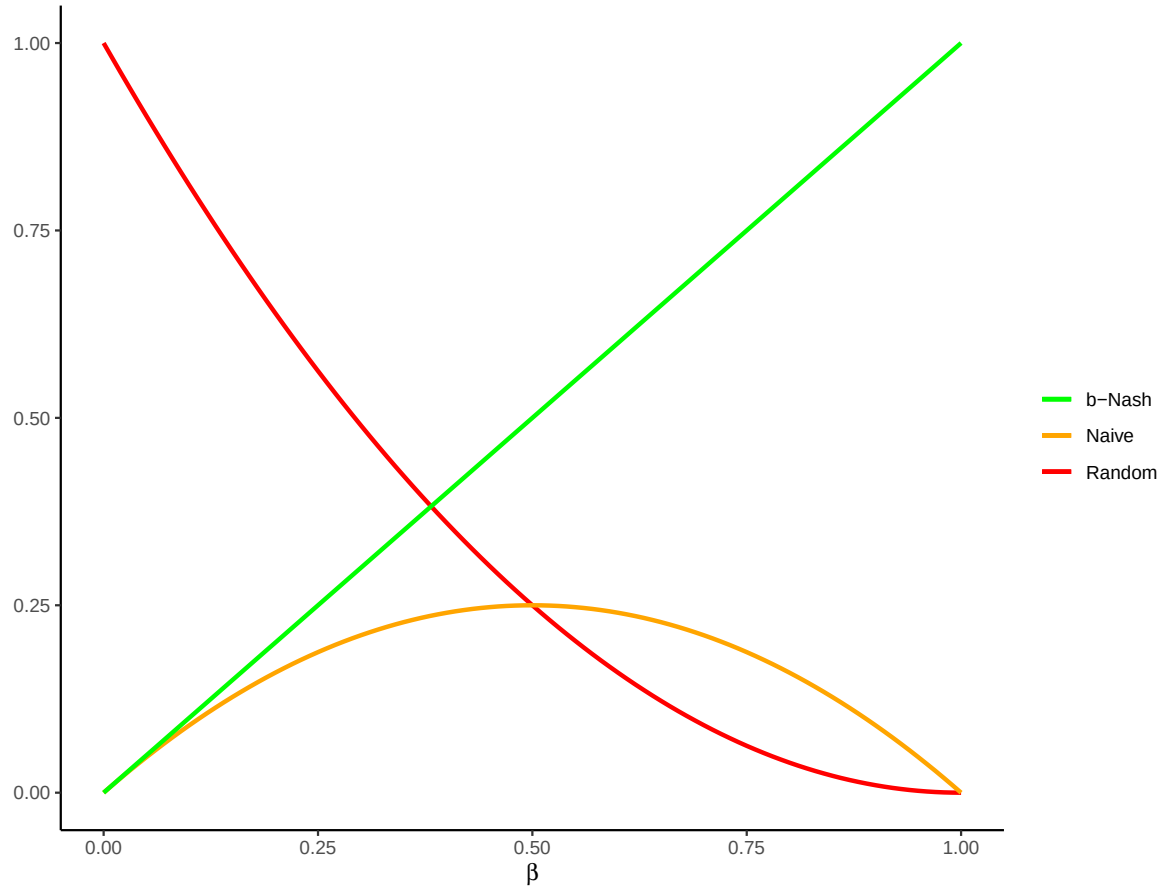


Figure 2.6: Repartition of subtypes in function of β

Remark 4: When β goes to 0, all agents play randomly. On the contrary when β goes to 1, all players play Nash equilibrium. It is a nice property of the model that is shared with other models: level-k models (when the parameter goes from 0 to $+\infty$) and Quantal Response Equilibrium models (when the parameter goes from $+\infty$ to 0).

6. Estimations

To empirically assess the relevance of our model, we use two datasets from two different behavioral game theory papers: Georganas et al. (2015) and Goeree et al. (2018). In the former subjects play 4 Undercutting games (UC) and in the latter 6 Money Request games (MR). We estimate our model in those 10 symmetric games. When estimating our model, if more than one distribution meet the required conditions (see Remark 3), we choose the distribution in which b-Nash players play symmetric b-nash strategy that maximize payoffs. We compare our model with the two main one-parameter models: the Level-K model with a Poisson distribution of levels (Level-K) and the logit Quantal Response Equilibrium (QRE) model. .

6.1. Undercutting (UC) Games

UC games are two-player symmetric games. Players have to choose a strategy corresponding to an integer between 1 and n. As shown in Figure 2.7 corresponding to UC1, players can "steal" ten units to their opponent by choosing an integer just below their opponent's choice. In each of the UC game, a number m (in UC1 m=4) undercuts all numbers between m+1 and n. If both players choose 1, they both earn 1 unit. Detail regarding UC2, UC3 and UC4 are in Appendix B (Figure 2.13, 2.14, 2.15)

	1	2	3	4	5	6	7
1	1 1	10 -10	0 0	0 0	0 0	0 0	-11 0
2	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0
3	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0
4	0 0	0 0	-10 10	0 0	10 -10	10 -10	10 -10
5	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
6	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
7	0 -11	0 0	0 0	-10 10	0 0	0 0	-11 -11

Figure 2.7: Undercutting Game 1 (UC1) as shown in Georganas et al. (2015) paper

In Figure 2.8 we decompose the distribution of strategies in UC1 predicted by the better beliefs model with β is equal to 0.582.[7]. With such β, 58.2% of players in the pool are strategic while 41.8% are not forming beliefs. 41.8% of those 41.8% (i.e around 17.5% of players in the pool) play randomly and therefore play with probability 1/7 each strategy (in red in Figure 2.8). 58.2% of 41.8% (i.e around 24.3% of players in the pool) play naively and choose to play 4 since it undercuts 5, 6 and 7 (in orange in Figure 2.8). Regarding the strategic players, they play diverse strategies depending on their belief b. For instance if b is close to 0 they play 4 and if b is close to 1 they play 1. Overall, they jointly play a mix strategy of 1, 2, 3 and 4 (in green in Figure 2.8).

[7] 0.582 is the value of the parameter estimated in UC1 with MSE estimation as we will see in the next subsection

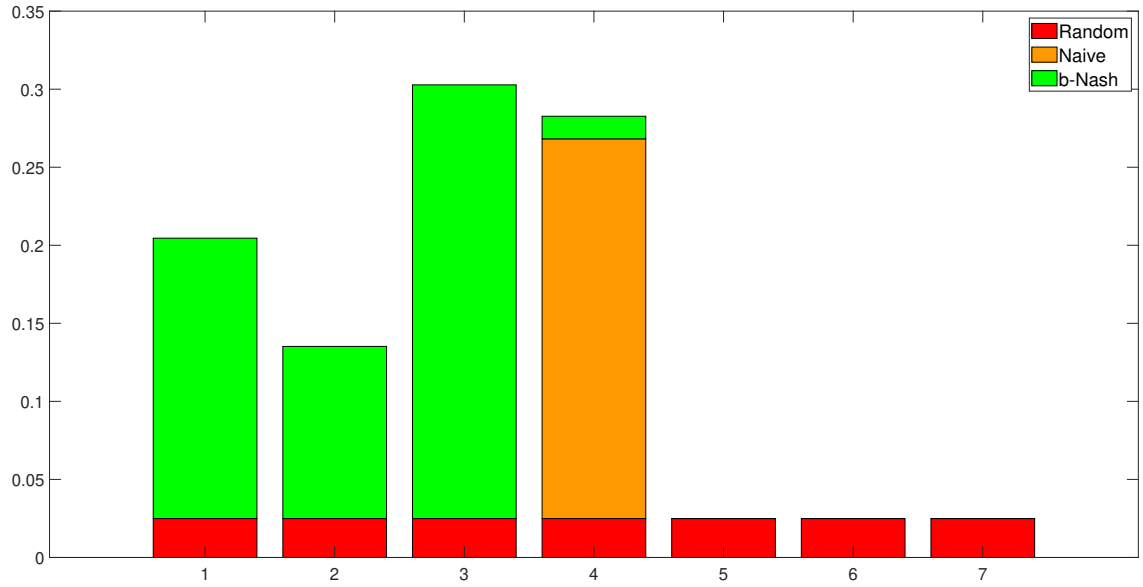


Figure 2.8: Distribution of subtypes in UC1

Figure 2.9 shows how the distribution of strategies predicted by our model as described above in UC1 fits the empirical distribution. We observe that our models predicts the four main empirically played strategies and also that 2 will be less played than 1, 3 and 4. While not perfect, our model fit more than correctly empirical behaviors.

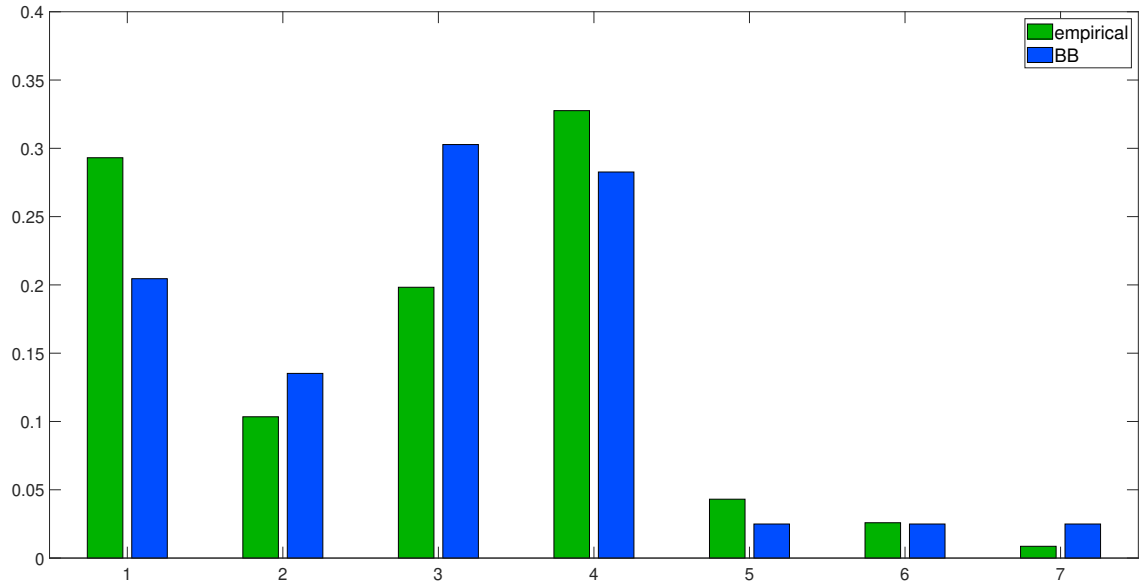


Figure 2.9: Distribution of empirical and predicted by BB model strategies in UC1

6.2. Money Request (MR) Games

MR games are two-player symmetric games. They are variations of the 11-20 Money Request game introduced by Arad and Rubinstein (2012). Players have to choose between ten boxes containing various amounts. They earn the amount of the box selected. In addition, if they select the box exactly one to the left of the box chosen by their opponent, they earn an additional amount R. Figure 2.10 shows the 6 MR games from Goeree et al. (2018). In treatment 11-20 (MR1, MR2 and MR3) R is equal to 20 while in treatment 1-10 (MR4, MR5 and MR6) R is equal to 8.

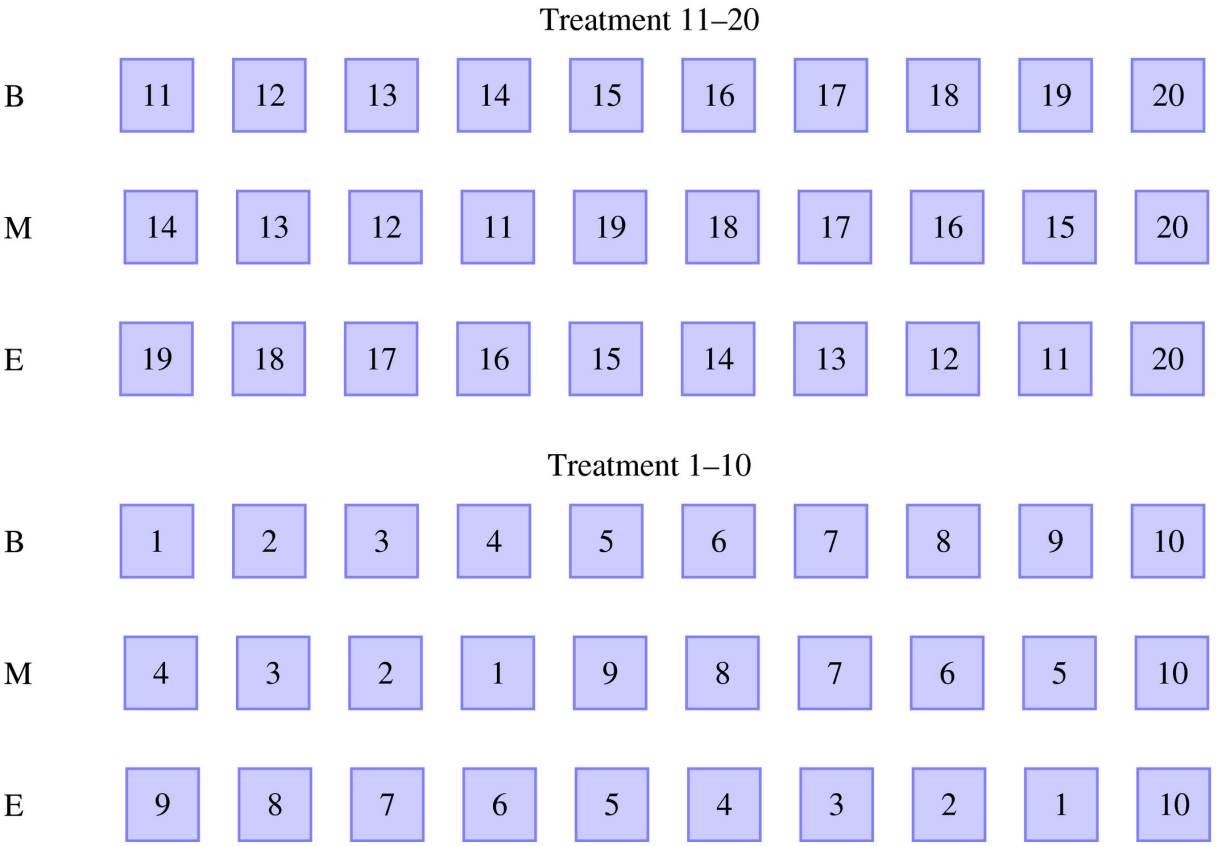


Figure 2.10: Money Request Games (MR1 to MR6) as shown in Goeree et al. (2018) paper

In Figure 2.11, we decompose the distribution of strategies in MR1 predicted by our model with β is equal to 0.495.[8]. While random players (in red) play randomly, naive players (in orange) choose 19. Overall, strategic players (in green) play a mix of 15, 16, 17, 18, 19 and 20.

[8] 0.495 is the value of the parameter estimated in UC1 with MSE estimation as we will see in the next subsection

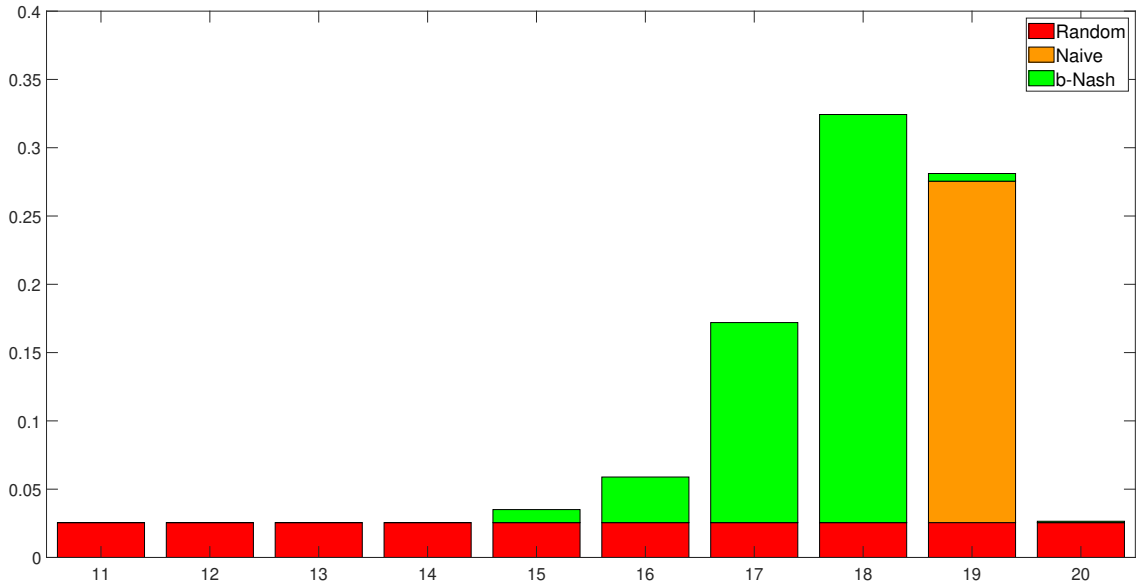


Figure 2.11: Distribution of subtypes in MR1

Figure 2.12 shows how the distribution of strategies predicted by our model as described above in MR1 fits the empirical distribution. Again, we observe that our models predicts the main empirically played strategies in particular the strategies 17, 18 and 19.

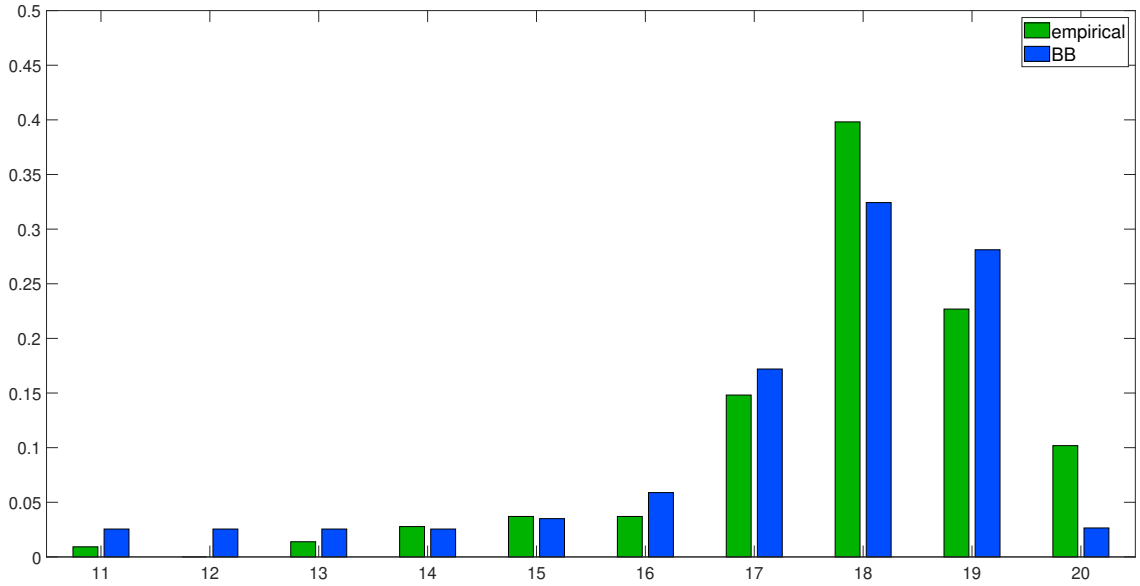


Figure 2.12: Distribution of empirical and predicted by BB model strategies in MR1

6.3. Best Estimation in each game

In this subsection, we estimate the best parameter for each model considered, for each and every game. We do this regarding two criteria: Maximum Likelihood and Mean

Squared Error. Detail regarding the estimations can be found in Appendix C. We report the obtained results in Table 2.2 and in Table 2.3.[9]

Table 2.2: ML Estimation in the ten games

	QRE		Level-K		BB	
Game	μ	LL	τ	LL	β	LL
UC1	2.16	-194.3	2.21	-191.4	0.600	**-186.6**
UC2	1.61	-199.4	2.23	-194.4	0.633	**-187.9**
UC3	1.27	-222.9	0.34	-240.3	0.764	**-212.7**
UC4	2.40	-193.9	2.19	-183.4	0.654	**-174.3**
MR1	3.10	-386.3	1.54	-395.5	0.459	**-384.6**
MR2	4.27	**-461.3**	0.16	-480.8	0.459	-461.4
MR3	3.45	-435.9	1.20	**-403.7**	0.408	-406.6
MR4	1.93	-354.5	2.36	-362.1	0.753	**-349.2**
MR5	1.75	**-385.7**	0.50	-438.6	0.177	-444.7
MR6	1.20	**-292.3**	0.52	-404.0	0.228	-404.0

Table 2.3: MSE Estimation in the ten games

	QRE		Level-K		BB	
Game	μ	MSE	τ	MSE	β	MSE
UC1	2.35	44.05	2.32	47.96	0.582	**31.98**
UC2	1.77	36.48	2.26	41.34	0.614	**21.18**
UC3	1.28	67.36	0.28	96.03	0.801	**32.67**
UC4	2.32	63.01	2.17	55.30	0.654	**16.30**
MR1	3.19	41.68	1.94	30.83	0.495	**16.18**
MR2	4.36	18.59	0.18	31.38	0.670	**18.24**
MR3	3.13	33.53	1.45	**8.61**	0.425	12.32
MR4	0.78	10.70	2.59	42.20	0.834	**3.68**
MR5	1.18	**12.99**	0.49	55.24	0.162	60.66
MR6	1.09	**1.66**	0.52	102.94	0.228	102.94

We observe that our model correctly fits the data compared to the logit QRE and Poisson Level-K models. In all UC games, the Better Beliefs model outperforms both models, regardless of the estimation method. For MR games, results are more ambiguous

[9]For the Logit-QRE model, we adopt the same convention as in Goeree et al. (2018) with μ being the precision parameter $\mu = \frac{1}{\lambda}$ with λ the parameter used in original QRE paper by McKelvey and Palfrey (1995)

but the Better Beliefs model is often well ranked. Remember that β represents the proportion of players that is strategic. Here, most of the time, more than half of the subjects appear as strategic and less than half as confused, not forming beliefs. It is consistent with the literature assessing the proportion of confused subjects in lab experiments.

Remark 4: MR4, MR5 and MR6 are unusual games and strongly disadvantage level-k models as it is shown in Goeree et al. (2018). The reason is the bonus being only 8. Thus, the best response to random players is to play 10 in all of those three games. In MR6 it is particularly striking since except random players, all level-k players are supposed to play 10. The same thing happens in our model since in MR6, all but random players are supposed to play 10 with probability 1. In those games, the *better response* attribute of players seems to explain better behavioral strategies. However, the *better-beliefs* attribute of players seems to explain better behaviors in UC games. In Appendix E, we introduce a variant of our model in which strategic players better-respond to their better beliefs. We find that allowing strategic players to tremble when best-responding make the better beliefs account for the behaviors observed in those three games.

6.4. Out of Sample Predictions

Fitting data ex-post is important. But a model also needs to have some predictive power. To compare models, we here fit the single parameter of the three models in UC1 (respectively MR1) and then compute how competitive they are in the other UC games with this very parameter (respectively other MR games). Again we do this using Maximum Likelihood Estimation and Mean Squared Error Estimation.

Table 2.4: Log-likelihood of the distribution of predicted strategies

	QRE		Level-K		BB	
Game	μ	LL	τ	LL	β	LL
UC2	2.16	-201.4	2.21	-194.4	0.600	**-188.2**
UC3	2.16	-226.5	2.21	-265.3	0.600	**-223.0**
UC4	2.16	-194.0	2.21	-183.4	0.600	**-175.1**
MR2	3.10	-465.1	1.54	-620.0	0.459	**-461.4**
MR3	3.10	-436.7	1.54	**-407.9**	0.459	**-407.9**
MR4	3.10	**-370.3**	1.54	-387.2	0.459	-391.6
MR5	3.10	**-398.4**	1.54	-480.3	0.459	-478.2
MR6	3.10	**-342.7**	1.54	-467.3	0.459	-440.1

Table 2.5: Mean Squared Error of the distributions of predicted strategies

	QRE		Level-K		BB	
Game	μ	MSE	τ	MSE	β	MSE
UC2	2.35	41.19	2.32	41.48	0.582	**21.83**
UC3	2.35	**77.57**	2.32	148.36	0.582	79.56
UC4	2.35	63.02	2.32	56.62	0.582	**21.92**
MR2	3.19	**22.44**	1.94	151.23	0.495	26.25
MR3	3.19	33.57	1.94	16.71	0.495	**14.96**
MR4	3.19	**52.02**	1.94	66.61	0.495	77.82
MR5	3.19	**36.14**	1.94	105.75	0.495	154.42
MR6	3.19	**81.92**	1.94	287.27	0.495	207.87

We observe that in UC games, except for UC3 with MSE criteria in which it is slightly worse than QRE, our model outperforms both other models in terms of predictive power. In MR games, results are more contrasted but except for games tackled in Remark 4, our model does correctly predict empirical strategies.

7. Conclusion

The better beliefs model has nice properties and its theoretical existence is ensured by semi-algebraic geometry arguments. Furthermore, our one-parameter model correctly explains and predicts the data from two behavioral game theory papers. We introduce noise in beliefs micro-founded by the approximate estimation of the probability to face a not forming beliefs players by strategic players.

To account for the heterogeneity of plays in experimental games, behavioral models need to include some noise. While the QRE model introduces noise in the best-response process, we introduce noise in the belief formation process. While our model outperforms QRE in some games, demonstrating the usefulness of *better beliefs*, some experimental data cannot correctly be explained without *better response*. In particular, if the cost of a mistake is mild, some strategies may be played with significant probabilities in the QRE. In contrast, the better beliefs model, because it assumes perfect best-response, relies on an *ordinal* structure: a strategy that yields a lower payoff than another will never be chosen, even if the difference is insignificant. Think, for instance, of a strategy that provides a payoff of 10 and one that provides a payoff of 9.99. The *ordinal* property of best-responses implies that the strategy with a payoff of 10 will *always* be preferred.

References

Agranov, M., A. Caplin, and C. Tergiman (2015). Naive play and the process of choice in guessing games. *Journal of the Economic Science Association 1*(2), 146–157.

Agranov, M., E. Potamites, A. Schotter, and C. Tergiman (2012). Beliefs and endogenous cognitive levels: An experimental study. *Games and Economic Behavior 75*(2), 449–463.

Alaoui, L. and A. Penta (2016). Endogenous depth of reasoning. *The Review of Economic Studies 83*(4), 1297–1333.

Arad, A. and A. Rubinstein (2012). Multi-dimensional iterative reasoning in action: The case of the colonel blotto game. *Journal of Economic Behavior & Organization 84*(2), 571–585.

Burchardi, K. B. and S. P. Penczynski (2014). Out of your mind: Eliciting individual reasoning in one shot games. *Games and Economic Behavior 84*, 39–57.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics 119*(3), 861–898.

Costa-Gomes, M. A. and G. Weizsäcker (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies 75*(3), 729–762.

De Sousa, J., G. Hollard, and A. Terracol (2013). Non-strategic players are the rule rather than the exception. Technical report, working paper, Université Paris 1.

Fragiadakis, D. E., D. T. Knoepfle, and M. Niederle (2016). Who is strategic? Technical report, Working Paper. 1.1.

Fragiadakis, D. E., A. Kovaliukaite, and D. R. Arjona (2019). Belief-formation in games of initial play: an experimental investigation.

Friedman, E. (2018). Stochastic equilibria: Noise in actions or beliefs? *Available at SSRN 3112977*.

Friedman, E. and J. Ward (2019). Stochastic choice and noisy beliefs in games: an experiment . Technical report, Mimeo.

Friedman, J. W. and C. Mezzetti (2005). Random belief equilibrium in normal form games. *Games and Economic Behavior 51*(2), 296–323.

Georganas, S., P. J. Healy, and R. A. Weber (2015). On the persistence of strategic sophistication. *Journal of Economic Theory 159*, 369–400.

Goeree, J. K. and C. A. Holt (2001). Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review 91*(5), 1402–1422.

Goeree, J. K. and C. A. Holt (2004). A model of noisy introspection. *Games and Economic Behavior 46*(2), 365–382.

Goeree, J. K. and P. Louis (2021). M equilibrium: A theory of beliefs and choices in games. *Available at SSRN 3829622*.

Goeree, J. K., P. Louis, and J. Zhang (2018). Noisy introspection in the 11–20 game. *The Economic Journal 128*(611), 1509–1530.

Hollard, G. and F. Perez (2020). Self-selection filters irrationality in one-shot games. Technical report, Center for Research in Economics and Statistics.

Koriyama, Y. and A. I. Ozkes (2021). Inclusive cognitive hierarchy. *Journal of Economic Behavior & Organization 186*, 458–480.

Levin, D. and L. Zhang (2019). Bridging level-k to nash equilibrium. *Available at SSRN 2934696*.

McKelvey, R. D. and T. R. Palfrey (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior 10*(1), 6 – 38.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review 85*(5), 1313–1326.

Nagel, R. et al. (1993). Experimental results on interactive competitive guessing. Technical report, University of Bonn, Germany.

Nash, J. F. et al. (1950). Equilibrium points in n-person games. *Proceedings of the national academy of sciences 36*(1), 48–49.

Rogers, B. W., T. R. Palfrey, and C. F. Camerer (2009). Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory 144*(4), 1440–1467.

Schanuel, S. H., L. K. Simon, and W. R. Zame (1991). The algebraic geometry of games and the tracing procedure. In *Game Equilibrium Models II*, pp. 9–43. Springer.

Stahl, D. and P. Wilson (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior 10*(1), 218–254.

Stahl, D. O. and P. W. Wilson (1994). Experimental evidence on players' models of other players. *Journal of economic behavior & organization 25*(3), 309–327.

Weizsäcker, G. (2003). Ignoring the rationality of others: evidence from experimental normal-form games. *Games and Economic Behavior 44*(1), 145–171.

Appendix 2.A Proofs

Proof of Proposition 1

Let Γ be a normal form game with the notations introduced in Section 2.

If $b = 0$, the problem faced by the strategic player is not a fixed-point problem but just a maximization problem that has at least one solution due to the compactness of X and continuity of u (u is multilinear in finite dimension hence continuous).

Fix $b \in]0;1]$ and $\sigma_N(b) \in X$.

Let $bBR^i : X^{-i} \to X^i$ such that $x^{-i} \to BR^i((1-b)\,\sigma_N^{-i}(b) \;+\; b\,x^{-i})$

Let $bBR : X \to X$ such that $x \to BR((1-b)\,\sigma_N(b) \;+\; b\,x)$

We need to show that bBR has a fix point. Since X is a non-empty, compact and convex subset of $\prod_i \mathbb{R}^{k_i}$, we will prove that bBR satisfies the conditions of Kakutani Theorem, namely: (1) $\forall x \in X$ $bBR(x)$ is non-empty and convex (2) bBR has a closed graph.

(1) Let $x \in X$, $i \in [\![1;n]\!]$ and $g^i : X^i \to \mathbb{R}$ such that $g^i(z) = u^i(z, (1-b)\sigma_N^{-i}(b) + b\,x^{-i})$. g^i is linear and therefore continuous (since X^i has a finite dimension) on the compact X^i thus reaches its maximum $M^i = max_{z \in X^i} g^i(z)$. Therefore $bBR^i(x^{-i})$ is non-empty. As g^i is linear $bBR^i(x^{-i})$ is convex. As it is the case for all $i \in [\![1;n]\!]$, $bBR(x)$ is non-empty and convex.

(2) Let $(x_n, y_n)_{n\in\mathbb{N}}$ a sequence in $Gr(bBR)$ with $(x,y) = \lim_{n\to+\infty}(x_n, y_n)$ then by definition: $\forall n \in \mathbb{N}, \forall i \in [\![1;n]\!], \forall z^i \in X^i, u^i(y_n^i, (1-b)\,\sigma_N^{-i}(b) + b\,x_n^{-i}) \geq u^i(z^i, (1-b)\,\sigma_N^{-i}(b) + b\,x_n^{-i})$. Taking the limit and using the continuity of u^i, we have $u^i(y^i, (1-b)\,\sigma_N^{-i}(b) + b\,x^{-i}+) \geq u^i(z^i, (1-b)\,\sigma_N^{-i}(b) + b\,x^{-i})$ and therefore $(x,y) \in Gr(bBR)$. Hence the closeness of Gr(bBR). QED

Proof of Proposition 2

Let Γ be a normal form game with the notations introduced in section 2. We will prove that the set-valued function $F : [0,1] \to X$

$$F : b \to \sigma_b \;\text{ s.t. }\; \sigma_b \in BR[\,(1-b)\,(b\sigma_{naive} + (1-b)\sigma_{rand}) \;+\; b\,\sigma_b)\,]$$

is semi-algebraic. Since $[0,1]$ and X are semi-algebraic, we just need to prove that Gr(F) is semi-algebraic. Let V_i be the following set:

$$V_i = \bigcup_{s_k^i \in S^i} \{(b,x) \in [0,1] \times X, \sum_{j=1}^{k_i} [bu^i(s_j^i, x^{-i}) + (1-b)bu^i(s_j^i, \sigma_{naiv}^{-i}) + (1-b)^2 u^i(s_j^i, \sigma_{rand}^{-i})]x_j^i$$
$$< bu^i(s_k^i, x^{-i}) + (1-b)bu^i(s_k^i, \sigma_{naiv}^{-i}) + (1-b)^2 u^i(s_k^i, \sigma_{rand}^{-i})\}$$

V_i is the set of (b,strategies) such that a pure strategy is strictly better for the player i in the fix point problem faced by the b-nash players.

Since u_i is multilinear and since $[0,1] \times S$ is semi-algebraic V_i is only defined by polynomial inequalities and is therefore a semi-algebraic set. Yet, we have that :

$$Gr(F) = \{[0,1] \times X\} \setminus \bigcup_{i=1}^{n} V_i$$

Therefore, Gr(F) is semi-algebraic.

Since for all $b \in [0,1]$ $F(b)$ is non-empty (by Proposition 1) we can apply the "Selections" proposition from Schanuel et al. (1991) to achieve the proof.

Appendix 2.B Games studied in the paper

Below you can find the description of UC2, UC3 and UC4 from Georganas et al. (2015)

	1	2	3	4	5	6	7	8	9
1	1 1	10 -10	0 0	0 0	0 0	0 0	0 0	0 0	-11 0
2	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0	0 0	0 0
3	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0	0 0
4	0 0	0 0	-10 10	0 0	10 -10	10 -10	10 -10	10 -10	10 -10
5	0 0	0 0	0 0	-10 10	0 0	0 0	0 0	0 0	0 0
6	0 0	0 0	0 0	-10 10	0 0	0 0	0 0	0 0	0 0
7	0 0	0 0	0 0	-10 10	0 0	0 0	0 0	0 0	0 0
8	0 0	0 0	0 0	-10 10	0 0	0 0	0 0	0 0	0 0
9	0 -11	0 0	0 0	-10 10	0 0	0 0	0 0	0 0	-11 -11

Figure 2.13: Undercutting Game 2 (UC2) as shown in Georganas et al. (2015) paper

	1	2	3	4	5	6	7	8	9
1	1 1	10 -10	0 0	0 0	0 0	0 0	0 0	0 0	-11 0
2	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0	0 0	0 0
3	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0	0 0
4	0 0	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0
5	0 0	0 0	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0
6	0 0	0 0	0 0	0 0	-10 10	0 0	10 -10	10 -10	10 -10
7	0 0	0 0	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
8	0 0	0 0	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
9	0 -11	0 0	0 0	0 0	0 0	-10 10	0 0	0 0	-11 -11

Figure 2.14: Undercutting Game 3 (UC3) as shown in Georganas et al. (2015) paper

	1	2	3	4	5	6	7
1	1 1	10 -10	0 0	0 0	0 0	0 0	-11 0
2	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0
3	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0
4	0 0	0 0	-10 10	0 0	30 -30	10 -10	10 -10
5	0 0	0 0	0 0	-30 30	0 0	0 0	0 0
6	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
7	0 -11	0 0	0 0	-10 10	0 0	0 0	-11 -11

Figure 2.15: Undercutting Game 4 (UC4) as shown in Georganas et al. (2015) paper

Appendix 2.C Detail regarding estimations

We consider two-player symmetric games. Let $S = (s_1, ...s_j)^2$ be the strategy space. Let z^k be the strategy played by player $k \in [\![1, K]\!]$ (K=116 players for UC games, K=216 players for MR games). Aggregated strategies can be represented by a vector of length j (the number of strategies) : $(p(s_1), ..., p(s_j))$ where $p(s_i)$ is the proportion of players playing s_i (or the probability to play s_i). We denote $(p^M(s_1|\theta), ..., p^M(s_j|\theta))$ the vector of aggregated strategies predicted by the model M with parameter θ and $(\hat{p}(s_1), ..., \hat{p}(s_j))$ the vector of empirical aggregated strategies, $\hat{p}_i = \frac{\sum_{k=1}^{K} \mathbb{1}_{z^k=s_i}}{K}$.

Maximum Likelihood Estimation

Given the introduced notations, the likelihood function for a model M and a parameter θ is

$$L(\theta|z) = \prod_{k=1}^{K} p^M(z_k|\theta)$$

In general, we prefer to maximize the log Likelihood and the optimization program is then:

$$\max_{\theta} \sum_{k=1}^{K} log(p^M(z_k|\theta))$$

Mean Squared Error Estimation

Mean Squared Error estimation is directly linked with the minimization of the euclidean distance between aggregated predicted strategies and empirical aggregated strategies. Precisely, the optimization program is:

$$\min_{\theta} \sum_{i=1}^{j} [100(p^M(s_i|\theta) - \hat{p}(s_i))]^2$$

Numerical details

All optimization program were solved with the Matlab function *fminunc*. The Better beliefs model we introduce requires the computation of an integral. It is performed with the Matlab function *trapz* using 10^3 data points.

Appendix 2.D Variance of the distribution of the beliefs

We assume that for each strategic player i:

$$b_i = \beta + \epsilon_i \qquad \text{with } \epsilon_i \sim \underline{\mathcal{N}}(0, \nu^2)$$

where $\underline{\mathcal{N}}(0, \nu^2)$ is a truncated normal distribution so that the belief $b_i \in [0, 1]$ (and so $\epsilon_i \in [-\beta, 1 - \beta]$). We assume furthermore $\nu^2 = k\beta(1 - \beta)$. We chose to set $k = 1/4$. In the present subsection of the appendix, we show that estimations do not change much with $k = 1/2$ or $k = 1/8$ as can be seen from Table 2.6 and Table 2.7.

Table 2.6: ML Estimation in the ten games with 3 different variance parameter

	k=1/8		k=1/4		k=1/2	
Game	β	LL	β	LL	β	LL
UC1	0.608	-187.6	0.600	-186.6	0.596	-185.8
UC2	0.641	-188.9	0.633	-187.9	0.637	-187.0
UC3	0.772	-211.6	0.764	-212.7	0.749	-214.8
UC4	0.662	-175.0	0.654	-174.3	0.665	-173.8
MR1	0.437	-383.8	0.459	-384.6	0.488	-385.4
MR2	0.463	-459.6	0.459	-461.4	0.387	-462.5
MR3	0.405	-408.0	0.408	-406.6	0.402	-405.8
MR4	0.738	-347.8	0.753	-349.2	0.773	-352.7
MR5	0.162	-445.8	0.177	-444.7	0.191	-443.0
MR6	0.228	-404.0	0.228	-404.0	0.228	-404.0

Table 2.7: MSE Estimation in the ten games with 3 different variance parameters

	k=1/8		k=1/4		k=1/2	
Game	β	MSE	β	MSE	β	MSE
UC1	0.586	39.08	0.582	31.98	0.583	26.13
UC2	0.621	27.13	0.614	21.18	0.628	16.46
UC3	0.793	29.09	0.801	32.67	0.818	41.13
UC4	0.658	20.74	0.654	16.30	0.679	12.86
MR1	0.442	14.79	0.495	16.18	0.578	15.84
MR2	0.590	16.21	0.670	18.24	0.742	20.87
MR3	0.399	12.78	0.425	12.32	0.419	12.42
MR4	0.801	5.86	0.834	3.68	0.875	3.25
MR5	0.157	61.26	0.162	60.66	0.170	59.56
MR6	0.228	102.94	0.228	102.94	0.228	102.94

Appendix 2.E Better response to better beliefs

We can extend our model by allowing strategic players to better-respond to their better beliefs. We denote this model as Logit-Better-Beliefs (LBB).

Let's define the logistic quantal response function r with parameter $\mu > 0$: $X \to X$ such that $\forall i \in [\![1; n]\!]$ and $\forall j \in [\![1; k_i]\!]$:

$$r(\sigma)_{ij} = \frac{e^{u(s_j^i, \sigma^{-i})/\mu}}{\sum_{k=1}^{k_i} e^{u(s_k^i, \sigma^{-i})/\mu}}$$

A strategic player with beliefs b choose $\sigma(b)$ such that:

$$\sigma(b) = r[\,(1-b)\,(b\sigma_{naive} + (1-b)\sigma_{rand}) \;+\; b\,\sigma(b)\,]$$

Brouwer fixed-point theorem ensures the existence, for each $b \in [0, 1]$, of a strategy that verifies the fixed point equation above. Here we do not formally prove the existence of the integral of the strategies by strategic players. Nonetheless we numerically compute it in Matlab for the ten games so that we can estimate the parameter β. We fix μ equals to 1. so that only the model still have only one free parameter. Since we allow strategic players to have better belief we do not want our strategic players to have too inaccurate response. Therefore we fix μ not too large but not too small since we extend the model to allow for better response. In Table 2.8 we estimate the best parameter β in the LBB model for each game. We find that allowing for better response particularly solves the issue we had in the last two MR games. In eight of ten games, LBB model fits better

than QRE model.

Table 2.8: ML Estimation in the ten games

	QRE		LBB	
Game	μ	LL	β	LL
UC1	2.16	-194.3	0.635	**-184.5**
UC2	1.61	-199.4	0.681	**-187.1**
UC3	1.27	-222.9	0.787	**-216.4**
UC4	2.40	-193.9	0.644	**-175.9**
MR1	3.10	-386.3	0.522	**-381.7**
MR2	4.27	**-461.3**	0.149	-474.1
MR3	3.45	-435.9	0.364	**-423.3**
MR4	1.93	**-354.5**	0.828	-363.3
MR5	1.75	-385.7	0.672	**-385.2**
MR6	1.20	-292.3	0.854	**-290.1**

Regarding out of sample predictions, Table 2.9 shows that LBB model does better than QRE in 6 out of 8 games (in particular in the two last MR games).

Table 2.9: Log-likelihood of the distribution of predicted strategies

	QRE		LBB	
Game	μ	LL	β	LL
UC2	2.16	-201.4	0.635	**-187.4**
UC3	2.16	-226.5	0.635	**-219.0**
UC4	2.16	-194.0	0.635	**-175.9**
MR2	3.10	**-465.1**	0.522	-497.3
MR3	3.10	-436.7	0.522	**-430.0**
MR4	3.10	**-370.3**	0.522	-390.0
MR5	3.10	-398.4	0.522	**-388.2**
MR6	3.10	-342.7	0.522	**-306.2**

Chapter 3

How Serious is the Measurement-Error Problem in Risk-Aversion Tasks?

Note: This chapter is co-authored with Guillaume Hollard and Radu Vranceanu. I am first author of this paper that is accepted for publication in the Journal of Risk and Uncertainty. The present version of the paper is the accepted version.[1]

Abstract

This paper analyzes within-session test/retest data from four different tasks used to elicit risk attitudes. Maximum-likelihood and non-parametric estimations on 16 datasets reveal that, irrespective of the task, measurement error accounts for approximately 50% of the variance of the observed variable capturing risk attitudes. The consequences of this large noise element are evaluated by means of simulations. First, as predicted by theory, the coefficient on the risk measure in univariate OLS regressions is attenuated to approximately half of its true value, irrespective of the sample size. Second, the risk-attitude measure may spuriously appear to be insignificant, especially in small samples. Unlike the measurement error arising from within-individual variability, rounding has little influence on significance and biases. In the last part, we show that instrumental-variable estimation and the ORIV method, developed by Gillen et al. (2019), both of which require test/retest data, can eliminate the attenuation bias, but do not fully solve the insignificance problem in small samples. Increasing the number of observations to N=500 removes most of the insignificance issues.

Keywords: Experiments; Measurement error; Risk-aversion, Test/retest; ORIV; Sample size.

[1]For this chapter I am specifically grateful to an anonymous referee, Olivier Armantier, Gwen-Jiro Clochard, Paolo Crosetto, Tamas Csermely, Jules Depersin, Delphine Dubart, Uwe Dulleck, Antonio Filippin, Jonas Fooken, Nikolaos Georgantzís, Lucas Girard, Yannick Guyonvarch, Xavier d'Haultfoeuille, Nicolas Jacquemet, Alexander Rabas, Gerardo Sabater-Grande, Tara White, and participants at the 10th International Conference of the ASFEE 2019 in Toulouse and the ESA European meeting 2019 in Dijon for their suggestions and remarks that have helped to improve this work.

1. Introduction

Economists explain individual heterogeneity in observed behavior by appealing to a number of key individual characteristics, such as risk attitudes or time preferences. For many years the gold standard in experimental economics consisted in eliciting risk-aversion by means of incentivized experiments, where choices have material consequences (Schildberg-Hörisch, 2018). A common research practice is to elicit this kind of individual characteristic via an initial task, and then use the resulting figure as an explanatory variable in subsequent regressions.

One source of intellectual discomfort with this method is the substantial within-individual variability in these incentive-based measures. For example, the correlations between different measures of risk attitudes for the same individual are typically small, even when the same task is repeated within a short period of time (Csermely and Rabas, 2016; Dulleck et al., 2015).

From an econometric perspective, within-individual variability can be interpreted as measurement error (Hey et al., 2009), which has well-known negative consequences: in OLS regressions, the coefficient on the explanatory variable that is measured with error is attenuated and, in multivariate regressions, other variables may falsely appear as significant, as the measurement error in one explanatory variable renders all of the estimates inconsistent (Pischke, 2007).

Another difficulty stems from the fact that a majority of popular elicitation methods yield a discrete approximation of a *continuous* variable (e.g. risk-aversion or the discount rate). Rounding elicited measures will mechanically generate some imprecision.

Last, the risk-aversion estimated in laboratory experiments often comes from relatively small samples, in particular in between-subject designs (e.g., N=100 or 200). Small samples will only amplify the measurement-error problem, as the variance of the estimated coefficients will be larger.

Measurement error, coupled with small sample sizes, raises questions regarding the robustness of the econometric analyses such as: What is the degree of attenuation of the coefficients in OLS regressions? and How often will significant coefficients actually appear to be insignificant? Furthermore, elicitation methods may differ in their test/retest stability; Is there a method that stands because of its low measurement error as compared to other methods?

Our analysis here proceeds in four steps. We first provide estimates of the extent of measurement error using both parametric (maximum-likelihood, ML) and non-parametric (NP) estimation methods, for 16 test/retest datasets covering four different risk-elicitation tasks. In a second step, we compare the size of the measurement error across the samples and methods.

The third step consists of the simulation a large number (100 000) of times of a

univariate linear stochastic model. We carry out OLS regressions with the independent variable being either the "true" risk-attitude measure, or noisy and/or rounded measures, over a variety of sample sizes. The simulations are calibrated using the parameters of the distributions as determined in the second step.[1] This allows us to disentangle the impact of measurement error and rounding on the size and significance of the estimated coefficient. In the last step, the simulations allow us to analyze and compare potential remedies for measurement error, such as increasing the number of observations, IV estimation, or using the Obviously Related Instrumental Variables (ORIV) method developed by Gillen et al. (2019).

In summary, we find that:

(1) Somewhat surprisingly, the four elicitation tasks considered, and the different datasets, generate similar levels of noise, as measured by the ratio of the variance of the error term to the variance of the observed risk-aversion measure. This result is robust to different estimation methods: in both maximum-likelihood (ML) and non-parametric estimations the variance of the measurement error is similar to that of the latent risk-aversion variable in all 16 datasets. The difference between the parametric and non-parametric estimates are only small, suggesting that the normality assumptions involved in the ML estimates (and neglecting the rounding effect in the non-parametric estimations) play only a marginal role in the results.

(2) Our simulations show that the discrete transformation of the variable of interest (i.e. rounding) affects the attenuation bias and the variance of the estimators only little. By way of contrast, the measurement error arising from within-subject variability is responsible for much of the attenuation effect.

(3) The attenuation factor is approximately 0.5 in all four of the elicitation methods considered. In line with theory, this holds *regardless of the size of the sample*. Our subsequent simulations confirm that the typical amount of noise in the risk-elicitation task divides the estimated coefficient on the variable of interest by around 2.

(4) Small sample sizes (e.g. $N = 100$ or $N = 200$) produce a large proportion of (falsely) insignificant coefficients at the standard significance levels. Increasing the sample size up to $N = 1000$ is sufficient for the coefficient to become significant almost every time. Intermediate values, such as N=500, already reduce the significance bias to a considerable extent.

(5) As expected, the ORIV method almost completely removes the attenuation bias, although the ORIV estimates do have larger variances than the true OLS estimates. ORIV may therefore not suffice to remove the significance issue resulting from measurement error in small samples.

Two contributions in the related literature have addressed the issue of measurement

[1] As a robustness check we also simulate a probit model.

error in experimental data. Gillen et al. (2019) replicate with a 6-month lag three classic risk experiments using an original dataset (the Caltech cohort survey), and show that the results can change dramatically when measurement error is correctly accounted for. Our analysis addresses two important elements that are not considered there: the impact of the sample size (in particular, the small sample size typical of laboratory experiments) and the rounding issue arising from the use of a discrete measure of a continuous variable. In addition, all test-retest data in our paper are collected within the same experimental session, which rules out any confounds affecting within-subject variability.

Engel and Kirchkamp (2019) adopt an alternative method to estimate the measurement error in the classical Holt and Laury (2002) task (or any multiple-price list tasks). Their analysis allows the error term to vary across each line, which may explain inconsistent answers,[2] which they use to estimate an individual-specific error term. In contrast, we here assume that the error terms are independent between the test and the retest, and are fixed within each task. We furthermore assume that error terms are drawn from the same distribution for all individuals. Under these assumptions, we can use the test/retest data to directly estimate the error variability. Our estimation strategy can be applied to any risk-elicitation task.

Many other contributions, as surveyed in Mata et al. (2018), used data collected with a substantial time lag between the test and the retest, spanning from several weeks to one year, and reveal a correlation that falls over time, in particular regarding incentivized tasks but also for survey-based measures (self-reported levels of risk aversion). In general, these results are interpreted as showing the evolution of preferences over the life cycle (Andersen et al., 2008; Lönnqvist et al., 2015; Bardsley et al., 2010; Beauchamp et al., 2017). To rule out this possible source of within-subject variability, we in this paper use only test/retest measures from the same session, in previously-published work.

The remainder of the paper is organized as follows. The next section describes the four elicitation tasks and the corresponding datasets. Section 3 introduces the parametric and non-parametric estimation methods, which are then used to estimate the measurement error, jointly with the mean and variance of the variable of interest. Section 4 presents the simulations. Last section 5 concludes.

[2] Jacobson and Petrie (2009) record a large number of such mistakes in a different experiment, and argue that they can provide information about the true population distribution of the risk-aversion coefficient.

2. The four risk-aversion tasks

2.1. Data

Researchers in experimental and behavioral economics appeal to different incentivized tasks to measure individual risk aversion. As our empirical strategy requires test/retest data, we first surveyed the literature to identify relevant datasets. We imposed only two restrictions. First, as noted above, measurement error is neatly inferred only if the test and the retest are close together in time. We thus only selected test/retest data that were collected in the same experimental session. Second, the number of observations must be large enough for asymptotic estimation to make sense, and we therefore only include in the analysis datasets with N>50.

There are unfortunately only few analyses that fulfill these conditions (test/retest, within-session, N>50). An internet search, and exchanges with authors of the test/retest studies (to whom we are very grateful for their sharing of the data) allowed us to identify 16 datasets relating to four different risk-aversion tasks. Table 3.1 summarizes these contributions. The first column indicates the risk-elicitation task (as described in the next subsections), the second the paper that first introduced the task, and the last that with the test/retest experiment data.

Table 3.1: Tasks and datasets used to estimate measurement error

Task	Introduced in	N° subjects	Data from
HL	Holt and Laury (2002)	175	Holt and Laury (2002)
HL	Holt and Laury (2002)	78	Dulleck et al. (2015)
AH1	Andreoni and Harbaugh (2009)	78	Dulleck et al. (2015)
AH2	Andreoni and Harbaugh (2009)	78	Dulleck et al. (2015)
AH3	Andreoni and Harbaugh (2009)	78	Dulleck et al. (2015)
AH4	Andreoni and Harbaugh (2009)	78	Dulleck et al. (2015)
AH5	Andreoni and Harbaugh (2009)	78	Dulleck et al. (2015)
AH6	Andreoni and Harbaugh (2009)	78	Dulleck et al. (2015)
AH7	Andreoni and Harbaugh (2009)	78	Dulleck et al. (2015)
AH8	Andreoni and Harbaugh (2009)	78	Dulleck et al. (2015)
AH9	Andreoni and Harbaugh (2009)	78	Dulleck et al. (2015)
SG1	Sabater-Grande and Georgantzis (2002)	208	García-Gallego et al. (2011)
SG2	Sabater-Grande and Georgantzis (2002)	208	García-Gallego et al. (2011)
SG3	Sabater-Grande and Georgantzis (2002)	208	García-Gallego et al. (2011)
SG4	Sabater-Grande and Georgantzis (2002)	208	García-Gallego et al. (2011)
BRET	Crosetto and Filippin (2013)	61	Crosetto and Filippin (2013)

The following subsection provides a definition and description of the variable of interest

in each of these tasks.

2.2. The Holt and Laury task (HL)

Holt and Laury (2002) (HL) is perhaps the most popular risk-aversion elicitation task in experimental economics; for the record, it had received over 6000 citations on Google Scholar as of March 21st 2021. [3] The (HL) risk-aversion elicitation task consists in choosing between a "safe" (small-spread) lottery $\frac{x}{10}.2\$ + (1 - \frac{x}{10}).1.6\$$ and a "risky" (wide-spread) lottery $\frac{x}{10}.3.85\$ + (1 - \frac{x}{10}).0.10\$$ for $x \in [\![1, 10]\!]$.

Table 3.2: The Holt and Laury (2002) risk-aversion elicitation task

Option A	Option B
1/10 of \$2.00, 9/10 of \$1.60	1/10 of \$3.85, 9/10 of \$0.10
2/10 of \$2.00, 8/10 of \$1.60	2/10 of \$3.85, 8/10 of \$0.10
3/10 of \$2.00, 7/10 of \$1.60	3/10 of \$3.85, 7/10 of \$0.10
4/10 of \$2.00, 6/10 of \$1.60	4/10 of \$3.85, 6/10 of \$0.10
5/10 of \$2.00, 5/10 of \$1.60	5/10 of \$3.85, 5/10 of \$0.10
6/10 of \$2.00, 4/10 of \$1.60	6/10 of \$3.85, 4/10 of \$0.10
7/10 of \$2.00, 3/10 of \$1.60	7/10 of \$3.85, 3/10 of \$0.10
8/10 of \$2.00, 2/10 of \$1.60	8/10 of \$3.85, 2/10 of \$0.10
9/10 of \$2.00, 1/10 of \$1.60	9/10 of \$3.85, 1/10 of \$0.10
10/10 of \$2.00, 0/10 of \$1.60	10/10 of \$3.85, 0/10 of \$0.10

Assuming that subjects maximize their expected utility,[4] and that their utility function is twice-differentiable, the value x^* (a continuous variable) for which the subject is indifferent between the safe (Option A) and the risky (Option B) lottery is strictly increasing in the coefficient of risk-aversion. x^* ($\in [0, 10]$) is thus a valid measure of risk preferences, and is our variable of interest for the HL measures. We only observe a discrete approximation to this measure, referring to the discrete number of safe choices (see Section 3). The retest data were collected at the end of the experiment in a "return to baseline" condition that replicated the first condition described in Table 3.2. In the meantime, subjects made four similar choices with different payoffs. The full set of data (175 observations) is provided by Holt and Laury (2002) in an online appendix. A second set of data (78 observations) was provided by Dulleck et al. (2015).

[3] This is acknowledged, for instance, in Zhou and Hey (2018), Charness et al. (2020), Attanasi et al. (2018) and Crosetto and Filippin (2016).

[4] The debates around this standard decision model are beyond the scope of the current paper; see O'Donoghue and Somerville (2018) for a recent discussion.

2.3. The Convex Risk Budget Task (AH)

The Convex Risk Budget Task (AH) is a risk-elicitation task introduced by Andreoni and Harbaugh (2009). At the onset of the experiment, a subject receives a budget b. Subjects have to choose a lottery, out of a set of simple binary lotteries with probability $x\%$ of winning a reward r, and a probability $(100 - x)\%$ of obtaining nothing. There is a mechanical relationship between x and r, such that larger rewards are less likely to be won: $r = b - xe$, with the key parameter being the "price", e, of increasing x by one percentage point. The individual thus chooses the couple (x, r). Various treatments can be considered with various values of b and e. For instance, consider a subject who receives a budget of \$100 and is facing a price $e = 2$. If he/she invests \$20, he/she will end-up facing a lottery with $x = 10\%$ and $r = \$80$. More risk-averse subjects will choose high-winning probability and low-prize lotteries, and risk-lovers high-prize and low winning-probability lotteries. x' Here the variable of interest $x^* \in [0, \frac{b}{e}]$ is the preferred wining probability, expressed in percentage points. We only observe a rounded value of x^*, as the value chosen by subjects is discrete.

The data for this task were also kindly shared with us by Dulleck et al. (2015). They carried out 9 different test/retest AH tasks in the same session. Table 3.3 presents the various values of b and e used in the nine tasks.

Table 3.3: The parameters of the nine AH tasks in the test-retest experiment of Dulleck et al. (2015)

	AH1	AH2	AH3	AH4	AH5	AH6	AH7	AH8	AH9
b	27.3	56	172	88	49.4	39.2	54.5	207	116
e	0.28	1.17	10.75	2.75	0.77	0.41	0.68	8.62	2.42

AH refers to Andreoni-Harbaugh task, for instance AH2 refers to the second Andreoni-Harbaugh task. b (respectively e) is the budget (respectively the cost of increasing the probability of winning of one percent) in the corresponding task.

2.4. The Lottery Choice Task (SG)

The Lottery Choice Task was introduced by Sabater-Grande and Georgantzis (2002). Subjects carry out four lottery choice tasks, each of which consists in choosing, within a given panel, a binary lottery with winning probability $\frac{x}{10}$ and reward r. In each panel the winning probability falls from 1 $(\frac{10}{10})$ (a sure rewards) to $\frac{1}{10}$, while at the same time the reward rises. Similar to the previous task, subjects face a trade-off between a higher reward x and a lower winning probability p. The payoffs and probabilities vary across the four panels from which subjects select a lottery. Compared to the convex risk budget task described above, this task involves a non-linear trade-off between risk and reward (see Table 3.4): the probability $\frac{x}{10}$ is associated with a reward $r = \frac{10 + t(10 - x)}{x}$, with

$t = 0.1$ (resp. 1, 5 and 10) for SG1 (resp. SG2, SG3 and SG4). The variable of interest $x^* \in [0, 10]$ reflects the subject's preferred probability in the task. Again, we have a rounded value of this preferred probability, as the value subjects choose is discrete.

Test/retest within session data were kindly offered to us by the authors of the task (García-Gallego et al., 2011).

Table 3.4: Four panels of ordered lotteries in García-Gallego et al. (2011)

SG1		SG2		SG3		SG4	
Prob.	Payoff	Prob.	Payoff	Prob.	Payoff	Prob.	Payoff
$\frac{1}{10}$	10.90€	$\frac{1}{10}$	19.00€	$\frac{1}{10}$	55.00€	$\frac{1}{10}$	100.00€
$\frac{2}{10}$	5.40€	$\frac{2}{10}$	9.00€	$\frac{2}{10}$	25.00€	$\frac{2}{10}$	45.00€
$\frac{3}{10}$	3.57€	$\frac{3}{10}$	5.70€	$\frac{3}{10}$	15.00€	$\frac{3}{10}$	26.70€
$\frac{4}{10}$	2.65€	$\frac{4}{10}$	4.00€	$\frac{4}{10}$	10.00€	$\frac{4}{10}$	17.50€
$\frac{5}{10}$	2.10€	$\frac{5}{10}$	3.00€	$\frac{5}{10}$	7.00€	$\frac{5}{10}$	12.00€
$\frac{6}{10}$	1.73€	$\frac{6}{10}$	2.30€	$\frac{6}{10}$	5.00€	$\frac{6}{10}$	8.30€
$\frac{7}{10}$	1.47€	$\frac{7}{10}$	1.90€	$\frac{7}{10}$	3.57€	$\frac{7}{10}$	5.70€
$\frac{8}{10}$	1.27€	$\frac{8}{10}$	1.50€	$\frac{8}{10}$	2.50€	$\frac{8}{10}$	3.80€
$\frac{9}{10}$	1.12€	$\frac{9}{10}$	1.20€	$\frac{9}{10}$	1.67€	$\frac{9}{10}$	2.20€
$\frac{10}{10}$	1.00€	$\frac{10}{10}$	1.00€	$\frac{10}{10}$	1.00€	$\frac{10}{10}$	1.00€

SG refers to Sabater-Grande and Georgantzis tasks, for instance SG2 is the second Sabater-Grande and Georgantzis task. Each task consists in choosing a row corresponding to a lottery with probability "Prob." of earning "Payoff" and probability 1-"Prob." of not earning anything.

2.5. The Bomb Risk Elicitation Task (BRET)

The Bomb Risk Elicitation Task (BRET) was developed by Crosetto and Filippin (2013). In the standard version of the task, subjects face a 10×10 matrix. Each cell represents a box. A subject can "collect" boxes one after the other. He/she can stop at any time, after collecting as many boxes as they wish. However, one random box in the matrix contains a (hidden) bomb, programmed to explode after the subject has made all of his/her choices. Let $x \in [\![0, 100]\!]$ be the number of collected boxes. If the subject does not collect the bomb, his/her dollar payoff is proportional to the number of boxes. More precisely he/she receives $\gamma * x$ dollars, where γ is the value of a box. However, if the bomb was in a collected box, the payoff is zero. The more boxes a subject collects, the higher is not only the potential payoff but also the risk of it vanishing. If the subject collects all 100 boxes, he/she must have collected the bomb and the payoff is zero for sure. Our parameter of interest is $x^* \in [0, 100]$ the possibly continuous preferred number of boxes collected. We only observe a rounding of this preferred number of boxes collected, as the

value chosen by subjects is discrete.

The test/retest within-session data were kindly provided by the authors of the task.

2.6. Latent and observed variables: A summary

As we can see, the four tasks are quite different in their implementation. The task in Holt and Laury (2002) is a standard Multiple Price List (MPL), the BRET task in Crosetto and Filippin (2013) is a sequential choice with risk accumulation, and the other two tasks involve the choice of a preferred lottery within a set of lotteries, with a variety of potential payoffs and winning probabilities. Table 3.5 summarizes the intervals of the latent continuous variable of interest and the discrete observed measure in each of the 16 elicitation tasks.

Table 3.5: Summary of the datasets: Intervals of the latent and actual values of the variable of interest

	Latent Variable of Interest x^*	Observed variable x
HL1	$[0, 10]$	$[\![0, 10]\!]$
HL2	$[0, 10]$	$[\![0, 9]\!]$
AH1	$[0, \frac{27.3}{0.28}]$	$[\![0, 97]\!]$
AH2	$[0, \frac{56}{1.17}]$	$[\![0, 47]\!]$
AH3	$[0, \frac{172}{10.75}]$	$[\![0, 16]\!]$
AH4	$[0, \frac{88}{2.75}]$	$[\![0, 32]\!]$
AH5	$[0, \frac{49.4}{0.77}]$	$[\![0, 64]\!]$
AH6	$[0, \frac{39.2}{0.41}]$	$[\![0, 95]\!]$
AH7	$[0, \frac{54.5}{0.68}]$	$[\![0, 80]\!]$
AH8	$[0, \frac{207}{8.62}]$	$[\![0, 24]\!]$
AH9	$[0, \frac{116}{2.42}]$	$[\![0, 47]\!]$
SG1	$[0, 10]$	$[\![1, 10]\!]$
SG2	$[0, 10]$	$[\![1, 10]\!]$
SG3	$[0, 10]$	$[\![1, 10]\!]$
SG4	$[0, 10]$	$[\![1, 10]\!]$
BRET	$[0, 100]$	$[\![0, 100]\!]$

The first column indicates the dataset. The two letters refer to a particular elicitation task (HL = Holt and Laury task, AH = Andreoni and Harbaugh task, SG = Sabater-Grande and Georgantzis task, and BRET = Bomb Risk Elicitation Task), and the number to a particular dataset.

3. Estimation strategies

3.1. Theory

The present section describes the two estimation strategies we use to gauge the magnitude of measurement error. The first is a parametric method using Maximum-Likelihood (ML) estimation, and the second is non-parametric (NP).

General Assumptions

For each risk-aversion task t of the 16 mentioned above, the variable of interest (i.e. the empirical measure of risk-aversion) is $x^* \in [0, M_t]$. For each task t, we observe two noisy measures of x^*, one during the test (first stage) and the other during the retest (second stage).

$$x'_1 = x^* + \epsilon_1 \quad \text{and} \quad x'_2 = x^* + \epsilon_2$$

with ϵ_1, ϵ_2 and x^* all being independent of each other. ϵ_1 and ϵ_2 are two zero-centered random variables with variance σ^2_ϵ, while x^* has a mean of m and a variance of σ^2_x. The main objective is to estimate σ^2_ϵ and σ^2_x to see whether the risk-aversion measures comprise a substantial amount of noise. In particular, we are interested in the ratio R, defined as the part of the measure's variance that reflects measurement error (noise):

$$R = \frac{\sigma^2_\epsilon}{\sigma^2_x + \sigma^2_\epsilon}$$

A low value of R suggests little measurement error. At the other extreme, a value close to 1 (or 100%) indicates that the elicited measure is composed almost only of noise.

As noted in the description of the tasks, one additional difficulty is that we do not observe x'_1 and x'_2 but rather the floor (for the HL measures) or the rounding (for the AH, SG and BRET measures) of x'_1 and x'_2, that is to say

$$x_i = \lfloor x'_i \rfloor \text{ (for HL)} \quad \text{or} \quad x_i = \lfloor x'_i + 0.5 \rfloor \text{ (for AH, SG and BRET)}$$

We will use both parametric and non-parametric methods to estimate the relevant variances and the ratio R.

The parametric method is based on maximum-likelihood estimation. The benefit of this method is that it takes into account rounding and truncating issues arising from the discrete elicitation of a continuous variable. The drawback, as in any parametric method, is that it requires specific assumptions regarding the distributions of the true risk-aversion parameter x^* and the measurement error ϵ.

The non-parametric approach, on the contrary, does not make any assumptions about the distribution of the measurement error and is easy to calculate. However, this crude measure cannot account for rounding and truncation.

As we will show later on in the results section, the two measures produce similar results, and together allow us to draw reliable conclusions about the extent of measurement error.

A Parametric Method: Maximum-Likelihood Estimation

Maximum-likelihood is a standard procedure to estimate the mean and variance of our variable of interest x^*, and the variance of the measurement error, ϵ_i.[5] To implement the ML method we introduce the following additional assumptions:

- The variable of interest is $x^* \sim \mathcal{N}(m, {\sigma_x}^2)$ truncated over $[0, M_t]$,[6] (with a density function of f).[7]
- The measurement error for observation $i \in \{1, 2\}$ is $\epsilon_i \sim \mathcal{N}(0, {\sigma_\epsilon}^2)$ (with a distribution function of Φ).
- ϵ_1, ϵ_2 and x^* are all independent of each other.
- We observe $x_1 = \lfloor x^* + \epsilon_1 \rfloor$ and $x_2 = \lfloor x^* + \epsilon_2 \rfloor$ (for the HL measures), or $x_1 = \lfloor x^* + \epsilon_1 + 0.5 \rfloor$ and $x_2 = \lfloor x^* + \epsilon_2 + 0.5 \rfloor$ (for the AH, SG and BRET measures)[8].
- The unknown parameters are $\theta = \{m, \sigma_x, \sigma_\epsilon\}$.

Under our assumptions, we can determine the likelihood function that measures, for any θ, the goodness of the model's fit to the sample of data $(x_{1i}, x_{2i})_{i \in [\![1,N]\!]}$.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N} P(x_1 = x_{1i} \bigcap x_2 = x_{2i} | \theta) \\ &= \prod_{i=1}^{N} \int_0^{M_t} P(x_1 = x_{1i} \bigcap x_2 = x_{2i} | x^* = u, \theta) f(u|\theta) du \end{aligned}$$

$$\widehat{\theta^{ML}} = (\widehat{m^{ML}}, \widehat{\sigma_x^{ML}}, \widehat{\sigma_\epsilon^{ML}}) = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$$

A Non-Parametric Method

Alternatively, measurement error can be estimated non-parametrically, which makes fewer restrictions on the data. We only assume independence between the errors ϵ and the true parameter x^*, and that the ϵ are independent and identically-distributed across repetitions.

Neglecting the rounding issue, we assume that $x_1 = x^* + \epsilon_1$ and that $x_2 = x^* + \epsilon_2$. As such, $Var(x_1 - x_2) = Var(\epsilon_1 - \epsilon_2) = 2Var(\epsilon)$, as ϵ_1 and ϵ_2 are assumed to be independent.

[5] The same method was used by Beauchamp et al. (2017) in a related analysis.
[6] σ_x^2 is the variance of x^* after truncation.
[7] In the Online Appendix C we discuss the plausibility of the normality assumption.
[8] In extreme cases we can observe 0 (resp M_t) if $x^* + \epsilon < 0$ (resp $x^* + \epsilon \leq \lfloor M_t \rfloor + 1$). We take this into account in the ML estimation.

$$Var(\epsilon) = \frac{Var(x_1 - x_2)}{2}$$

We can therefore estimate the variance of the measurement error using the empirical variances:

$$\widehat{\sigma_\epsilon^{NP}}^2 = \frac{\widehat{Var}(x_1 - x_2)}{2}$$

Then, using the same kind of reasoning,

$$\widehat{\sigma_x^{NP}}^2 = \frac{\widehat{Var}(x_1 + x_2) - 2\widehat{\sigma_\epsilon^{NP}}^2}{4}$$

For the HL measures:

$$\widehat{m^{NP}} = \widehat{\mathbb{E}}(\frac{x_1 + x_2}{2}) + 0.5$$

And for the AH, SG and BRET measures:

$$\widehat{m^{NP}} = \widehat{\mathbb{E}}(\frac{x_1 + x_2}{2})$$

3.2. Empirical estimates

This section presents the empirical estimates from the two methods above, i.e. the maximum-likelihood estimator $\widehat{\theta^{ML}} = \left\{\widehat{m^{ML}}, \widehat{\sigma_x^{ML}}, \widehat{\sigma_\epsilon^{ML}}\right\}$ and the non-parametric estimator $\widehat{\theta^{NP}} = \left\{\widehat{m^{NP}}, \widehat{\sigma_x^{NP}}, \widehat{\sigma_\epsilon^{NP}}\right\}$. For each dataset, the key variables of interest are the two ratios below, which are direct measures of the amount of noise generated by a particular risk-aversion task.

$$\widehat{R}^{ML} = \frac{\widehat{\sigma_\epsilon^{ML}}^2}{\widehat{\sigma_x^{ML}}^2 + \widehat{\sigma_\epsilon^{ML}}^2} \quad \text{and} \quad \widehat{R}^{NP} = \frac{\widehat{\sigma_\epsilon^{NP}}^2}{\widehat{\sigma_x^{NP}}^2 + \widehat{\sigma_\epsilon^{NP}}^2}$$

Each ratio is based on a particular estimation method: R^{ML} refers to the maximum-likelihood estimation described in the previous section and R^{NP} to non-parametric estimation.

As shown in Table 3.6, the values of R over the 16 tasks vary only little for a given estimation method (ML or NP). Furthermore, for any task, the difference between $R^{ML} - R^{NP}$ is fairly small, suggesting that the restrictive assumptions used for the calculation of the maximum-likelihood estimates play only a minor role.

Table 3.6: Key Estimates by Estimation Method and Risk Task

	Maximum Likelihood Estimation				Non-Parametric Method			
	$\widehat{m^{ML}}$	$\widehat{\sigma_x^{ML}}^2$	$\widehat{\sigma_\epsilon^{ML}}^2$	$\widehat{R}^{ML}$	$\widehat{m^{NP}}$	$\widehat{\sigma_x^{NP}}^2$	$\widehat{\sigma_\epsilon^{NP}}^2$	$\widehat{R}^{NP}$
HL1	5.72	1.06	0.809	**43.4%**	5.74	1.07	0.884	**45.3%**
HL2	5.96	1.68	1.31	**43.8%**	5.94	1.62	1.33	**45.2%**
AH1	46.4	116	107	**48.0%**	46.4	118	103	**46.5%**
AH2	27.4	30.3	33.0	**52.1%**	27.4	29.6	32.4	**52.3%**
AH3	9.83	5.75	3.36	**36.9%**	9.78	5.63	3.45	**38.0%**
AH4	19.4	14.6	20.1	**57.9%**	19.4	13.8	20.4	**59.6%**
AH5	35.1	54.8	73.5	**57.3%**	35.2	52.3	74.3	**58.7%**
AH6	48.1	122	134	**52.4%**	48.1	118	135	**53.3%**
AH7	44.5	84.9	95.5	**52.9%**	44.6	81.4	96.2	**54.2%**
AH8	13.9	17.3	9.86	**36.2%**	13.8	16.9	10.1	**37.4%**
AH9	28.2	59.0	40.3	**40.6%**	28.1	57.0	40.6	**41.6%**
SG1	2.96	3.19	1.78	**35.8%**	3.39	2.53	1.47	**36.7%**
SG2	3.69	1.22	1.28	**51.1%**	3.71	1.15	1.29	**53.0%**
SG3	4.15	0.958	1.03	**51.9%**	4.15	0.943	1.09	**53.7%**
SG4	3.96	1.00	1.39	**58.1%**	3.97	0.957	1.39	**59.2%**
BRET	43.4	149	173	**53.7%**	43.4	148	173	**53.9%**

The first column indicates the dataset. The two letters refer to a particular elicitation task (HL = Holt and Laury task, AH = Andreoni and Harbaugh task, SG = Sabater-Grande and Georgantzis task, and BRET = Bomb Risk Elicitation Task), and the number to a particular dataset..

Figure 3.1 summarizes the estimated ratios for all of the datasets, from both the ML and NP methods.

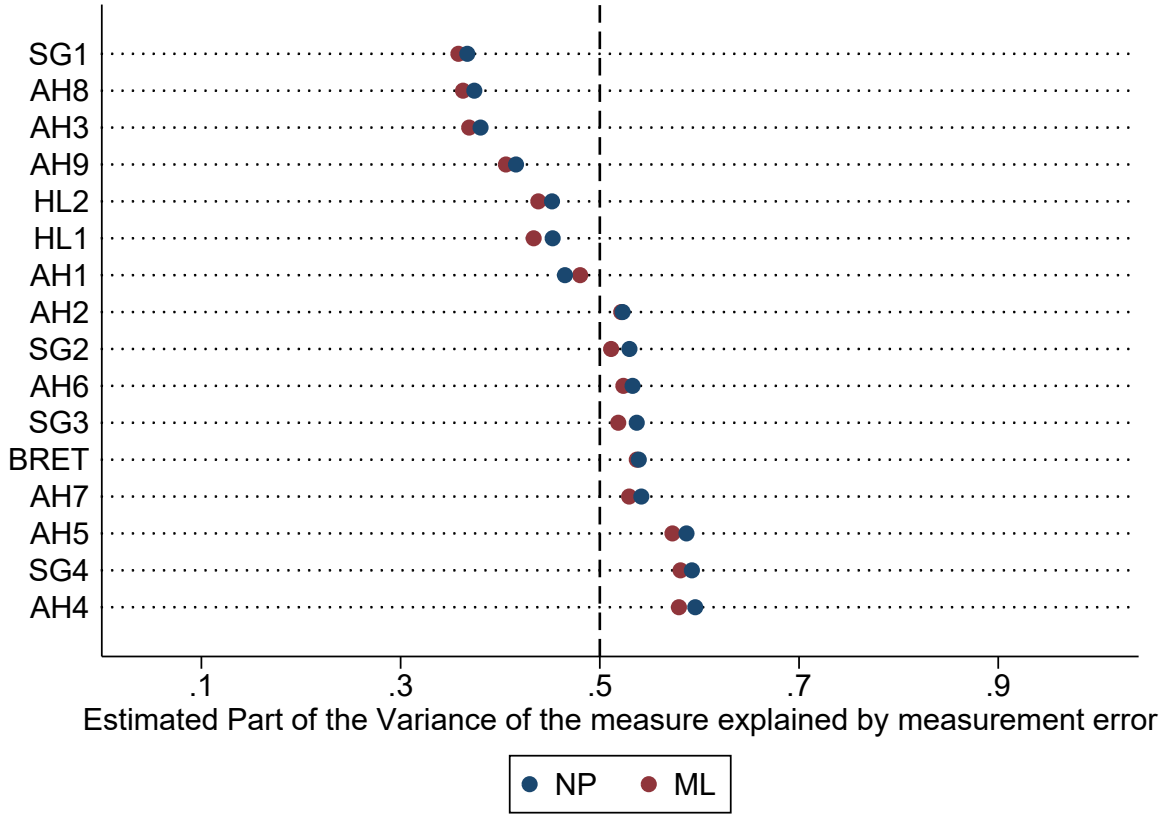


Figure 3.1: $\widehat{R}^{ML}$ and $\widehat{R}^{NP}$ for various tasks and datasets

The results allow us to draw three conclusions:

a) Based on the values of R, no risk-elicitation task emerges as being clearly better than the others. For example, the AH tasks appear both at the top of the figure (with lower values of R, and at the bottom with high values of R.

b) Using the same estimation method (ML or NP), the differences in the estimated Rs across the 16 tasks are only fairly small. As can be seen from Figure 3.1, most of the datasets yield estimates of R that are close to 0.5, ranging from 35% to 60%.

c) The estimated part of the variance that reflects measurement error is extremely similar when this is estimated by ML or NP.

It turns out that noise is a serious issue when measuring an individual's attitude towards risk. In the next section we analyze the consequences of this noise for classical regression analysis, and suggest ways of removing the ensuing biases.

4. Simulations

4.1. Fictive outcomes and assumptions

We first generate multiple datasets by means of a stochastic model, generating an outcome variable y^* that is linearly related to the variable of interest x^*. We then evaluate the size of the measurement-error problem in simple OLS regressions, focusing on the value, significance and variance of the estimated coefficient $\hat{\beta}$. The simulations are carried out under the following assumptions:

the true model is

- $y^* = \alpha + \beta x^* + u \qquad x^* \sim \mathcal{N}(m_X^*, {\sigma_X^*}^2) \qquad u \sim \mathcal{N}(0,1) \qquad X^* \perp\!\!\!\perp u$

and the observed variable is

- $x = \lfloor x^* + \epsilon \rceil \qquad$ with $\epsilon \sim \mathcal{N}(0, {\sigma_\epsilon}^2) \qquad X^* \perp\!\!\!\perp \epsilon \qquad \epsilon \perp\!\!\!\perp u$

4.2. Obviously Related Instrumental Variable

Gillen et al. (2019) argue convincingly that the test/retest design and the duplication of a noisy measure can help to correct attenuation bias and improve the significance of the estimated coefficients.

In a first step, they show that simple IV regressions (2SLS) using x_1 as an instrument for x_2 (or the reverse) already improve the quality of the estimation.

To make the best use of all available information, and because there is no reason to prefer x_1 to instrument x_2, or x_2 to instrument x_1, they combine the two IV regressions in one convex combination, via a method they called Obviously Related Instrumental Variables. This requires that the errors in the 1^{st} and 2^{nd} measures be independent.

We will implement both the IV and the ORIV methods, which will allow us to emphasize the benefits of the latter. For ORIV, we estimate the stacked model:

$$\begin{pmatrix} y^* \\ y^* \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \beta \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + u \tag{3.1}$$

$$\text{instrumenting} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{by } W = \begin{pmatrix} x_2 & 0_N \\ 0_N & x_1 \end{pmatrix}$$

4.3. The simulation outcomes

Table 3.7 lists the mean estimated coefficients from 100 000 simulated samples with N=100 subjects, a sample size that is relatively common in laboratory experiments. We use in the simulation five "actual" coefficients $\beta = (0.15, 0.20, 0.25, 0.30, 0.35)$, of a relatively small size, as these can be more sensitive to the measurement-error problem. The

parameters of the normal distributions of x^* and ϵ are those obtained using maximum-likelihood estimation based on the HL1 sample (the original study by Holt and Laury 2002), so that: $x^* \sim \mathcal{N}(\widehat{m^{ML}}, \widehat{\sigma_x^{ML}}^2)$ and $x^* \in [0, 10]$ and $\epsilon \sim \mathcal{N}(0, \widehat{\sigma_\epsilon^{ML}}^2)$. Using estimated values from other datasets will lead to very similar results, as the main driver of the attenuation is the R ratio previously defined (see Pischke, 2007, for the explicit calculation).

The table shows the estimated means, variances and frequencies of the significance of these estimators (at three significance levels). We stack the estimates by the method used to generate the latent variable (the true variable, discretization, noise, and last discretization and noise), and the estimation of the coefficient when the latent variable is noisy and truncated (by IV and ORIV). Given the values of the parameters used, we get Corr(y^*,x^*) $\simeq \beta$ so that coefficients are easy to interpret.[9]

To help intuition, Figure 3.2 depicts the distribution of the estimates for $\beta = 0.25$ in the panels of Table 3.7 for N=100; Table 3.8 displays the analogous estimates for a sample size of 200.

In Appendix A we provide coefficient estimates for "large" samples (up to N=1000), which appear much less frequently in laboratory experiments, but are common when using internet data collection through specialized platforms, or in some field studies. As expected, in these large samples the measurement-error problem regarding significance is much diminished.

[9]Precisely, for $\beta = \{0.15, 0.20, 0.25, 0.30, 0.35\}$, we get Corr($y^*$,$x^*$) $= \{0.157, 0.207, 0.256, 0.303, 0.347\}$

Table 3.7: Simulations: Simple OLS and IV with N=100 (100 000 simulations)

	β	0.15	0.2	0.25	0.3	0.35
x^*	Mean $\hat{\beta}$	0.1499	0.1999	0.2499	0.2999	0.3497
	(St Dev)	0.0985	0.0985	0.0985	0.0985	0.0987
	Sig 0.1	45.13%	64.85%	80.93%	91.38%	96.73%
	Sig 0.05	32.80%	52.26%	71.18%	85.29%	93.72%
	Sig 0.01	14.24%	28.54%	47.19%	66.49%	81.67%
$\lfloor x^* \rfloor$	Mean $\hat{\beta}$	0.1389	0.1853	0.2316	0.2780	0.3240
	(St Dev)	0.095	0.0950	0.0951	0.0952	0.0953
	Sig 0.1	42.73%	61.86%	78.21%	89.26%	95.65%
	Sig 0.05	30.85%	49.22%	67.78%	82.36%	91.85%
	Sig 0.01	12.96%	26.07%	43.26%	61.95%	77.75%
$x^* + \epsilon$	Mean $\hat{\beta}$	0.0851	0.1134	0.1417	0.1700	0.1979
	(St Dev)	0.0746	0.0749	0.0752	0.0757	0.0763
	Sig 0.1	30.96%	44.96%	59.40%	72.30%	82.32%
	Sig 0.05	20.48%	32.65%	46.79%	60.85%	72.92%
	Sig 0.01	7.39%	14.07%	23.78%	36.28%	49.67%
$\lfloor x^* + \epsilon \rfloor$	Mean $\hat{\beta}$	0.0814	0.1085	0.1356	0.1628	0.1894
	(St Dev)	0.0730	0.0733	0.0737	0.0742	0.0748
	Sig 0.1	30.01%	43.47%	57.82%	70.69%	80.77%
	Sig 0.05	19.90%	31.44%	44.94%	58.95%	71.07%
	Sig 0.01	7.03%	13.29%	22.50%	34.48%	47.27%
IV	Mean $\hat{\beta}$	0.1527	0.2035	0.2544	0.3053	0.3564
	(St Dev)	0.1416	0.1428	0.1443	0.1461	0.1487
	Sig 0.1	29.92%	43.75%	58.23%	71.36%	81.63%
	Sig 0.05	19.46%	31.11%	45.08%	59.50%	71.98%
	Sig 0.01	6.05%	12.17%	21.39%	33.37%	46.98%
ORIV	Mean $\hat{\beta}$	0.1527	0.2036	0.2544	0.3053	0.3561
	(St Dev)	0.1229	0.1240	0.1253	0.1270	0.1292
	Sig 0.1	37.18%	53.39%	68.61%	81.11%	89.56%
	Sig 0.05	26.11%	40.74%	57.01%	71.41%	82.76%
	Sig 0.01	10.71%	20.18%	33.15%	48.12%	62.96%

The first column indicates the variable or the estimation method used in the univariate OLS regression. The first variable is the true x^*, the second the discretization of the true variable, the third considers the effect of noise, the fourth combines noise and discretization. For the IV estimations, the discrete noisy measure $\lfloor x_1^* + \epsilon \rfloor$ is instrumented by $\lfloor x_2^* + \epsilon \rfloor$. For the ORIV estimations the stack model uses $\lfloor x_j^* + \epsilon \rfloor$ for $j \in \{1, 2\}$. The last five columns indicate the average value, standard deviation and significance of β for 100 000 simulations. For instance the ORIV cell for Sig 0.1 and $\beta = 0.15$ is 37.18%, so that the estimated β using ORIV is significant in 37.18% of the 100 000 regressions at the 10% level when the true β is 0.15.

Table 3.8: Simulations: Simple OLS and IV with N=200 (100 000 simulations)

	β	0.15	0.2	0.25	0.3	0.35
x^*	Mean $\hat{\beta}$	0.1499	0.1999	0.2499	0.2999	0.3499
	(St Dev)	0.0694	0.0694	0.0694	0.0694	0.0694
	Sig 0.1	69.79%	88.92%	97.26%	99.52%	99.95%
	Sig 0.05	57.94%	81.88%	94.63%	98.87%	99.85%
	Sig 0.01	33.74%	61.47%	84.03%	95.39%	99.04%
$\lfloor x^* \rfloor$	Mean $\hat{\beta}$	0.139	0.1853	0.2317	0.278	0.3243
	(St Dev)	0.0669	0.0669	0.0670	0.067	0.0671
	Sig 0.1	66.85%	86.87%	96.28%	99.26%	99.90%
	Sig 0.05	54.81%	78.93%	93.03%	98.33%	99.73%
	Sig 0.01	30.85%	57.35%	80.53%	93.63%	98.47%
$x^* + \epsilon$	Mean $\hat{\beta}$	0.0846	0.1129	0.1412	0.1695	0.1978
	(St Dev)	0.0524	0.0527	0.0529	0.0533	0.0536
	Sig 0.1	49.14%	69.14%	84.41%	93.55%	97.77%
	Sig 0.05	36.52%	57.47%	75.81%	88.60%	95.44%
	Sig 0.01	16.75%	33.13%	53.53%	72.14%	85.91%
$\lfloor x^* + \epsilon \rfloor$	Mean $\hat{\beta}$	0.0810	0.1081	0.1351	0.1622	0.1893
	(St Dev)	0.0513	0.0515	0.0518	0.0521	0.0525
	Sig 0.1	47.62%	67.50%	83.07%	92.64%	97.33%
	Sig 0.05	35.24%	55.48%	73.80%	87.14%	94.69%
	Sig 0.01	15.90%	31.50%	51.01%	69.68%	84.04%
IV	Mean $\hat{\beta}$	0.1512	0.2016	0.2520	0.3024	0.3528
	(St Dev)	0.0976	0.0984	0.0994	0.1006	0.1020
	Sig 0.1	48.13%	68.22%	83.50%	92.90%	97.40%
	Sig 0.05	35.50%	56.18%	74.63%	87.64%	94.93%
	Sig 0.01	14.49%	31.29%	51.39%	70.56%	84.78%
ORIV	Mean $\hat{\beta}$	0.1509	0.2013	0.2517	0.3021	0.3525
	(St Dev)	0.0851	0.0858	0.0866	0.0876	0.0888
	Sig 0.1	57.26%	77.76%	90.88%	97.06%	99.28%
	Sig 0.05	44.82%	67.58%	84.71%	94.30%	98.29%
	Sig 0.01	23.19%	43.78%	66.13%	83.41%	93.44%

The first column indicates the variable or estimation method used in the univariate OLS regression. The first variable is the true x^*, the second the discretization of the true variable, the third considers the effect of noise, and the fourth combines noise and discretization. For the IV estimations, the discrete noisy measure $\lfloor x_1^* + \epsilon \rfloor$ is instrumented by $\lfloor x_2^* + \epsilon \rfloor$. For the ORIV estimations the stack model uses $\lfloor x_j^* + \epsilon \rfloor$ for $j \in \{1, 2\}$. Columns 2 to 6 indicate the average value, standard deviation and significance of β for 100 000 simulations. For instance the ORIV cell for Sig 0.1 and $\beta = 0.15$ is 57.26%, so that the estimated β using ORIV method is significant in 57.26% of the 100 000 regressions at the 10% level when the true β is 0.15.

4.4. Main results

(a) In line with theory, the simulations confirm that measurement error attenuates the coefficient of the variable of interest in univariate OLS regressions: the "true" coefficient is approximately divided by 2. Increasing the size of the sample does not remove this bias, but does improve the significance of the estimated coefficients .

(b) In small samples (N=100), measurement errors substantially affect coefficient significance. For instance, with $\beta = 0.25$ the coefficient is significant at the 5% level only 46.79% of the time. This helps to explain why the coefficients of "meaningful" variables by any theoretical standard are often insignificant in experimental research .

(c) The use of a discrete measure of a continuous variable of interest such as risk-aversion does not appear to present a major problem. As we can see from the simulation tables, this transformation only slightly reinforces the downward bias in the coefficients.

(d) In small samples (N=100), simple IV and ORIV estimations do not fully remove the measurement problem: while the bias is virtually eliminated, the frequency of (falsely) insignificant coefficients is still high (at 55 and 43 percent, respectively, at the 5% significance level).

(e) In larger samples (N=200) the ORIV estimator performs relatively well. Not only is the bias virtually eliminated, but significance also improves (in particular as compared to the IV estimates). The ORIV coefficients are slightly upward-biased due to the discrete transformation of the observations. See Appendix B for a comparative performance analysis of IV and ORIV in large samples (N=1000).

The frequency curves in Figure 3.2 depict the distribution of the estimated coefficients (for $\beta = 0.25$, and N=100). These show that: (1) the main source of the bias is the measurement error; (2) the discrete transformation of the continuous variable of interest does not much shift the distribution; and (3) the ORIV method eliminates the bias but produces a higher variance.

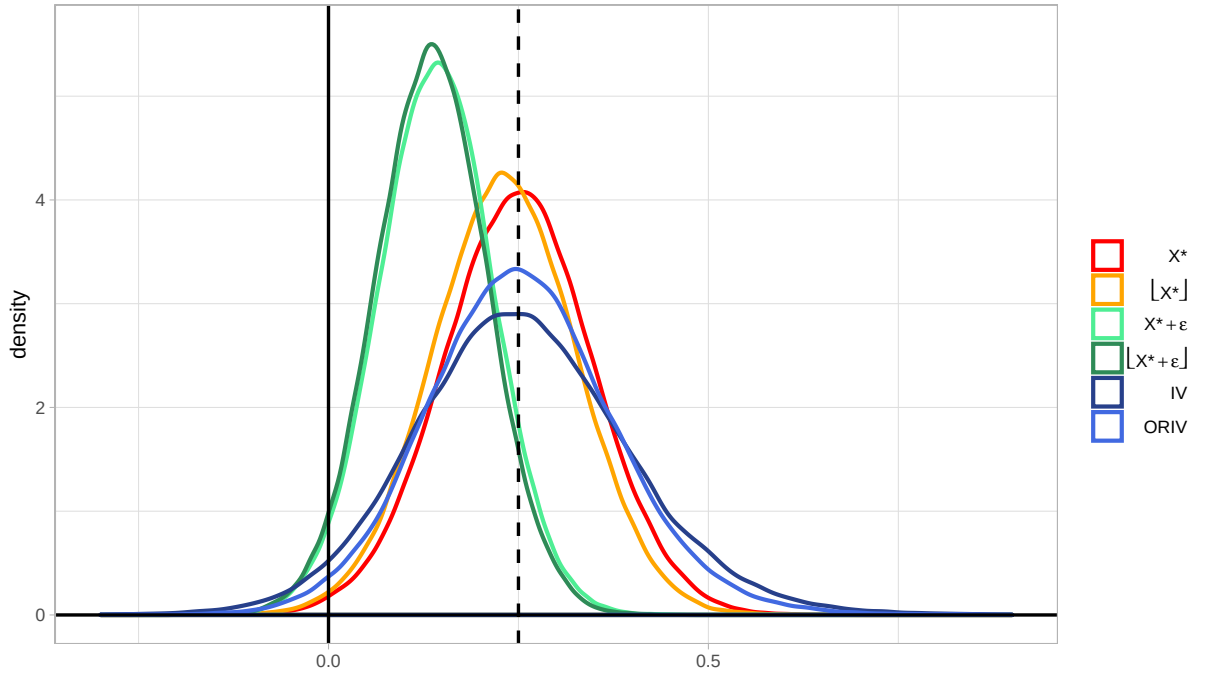


Figure 3.2: The distribution of the estimators for $\beta = 0.25$ and N=100

As a robustness check, we also performed simulations using Probit regressions to make sure that our results are not particular to OLS. The results presented in the Online Appendix B underline that our findings are not model-dependent.

5. Conclusion

In a recent paper, Gillen et al. (2019) pointed out that some of the standard measures used to elicit risk aversion and overconfidence might suffer from substantial measurement error. We have here extended their empirical analysis to test/retest data collected *within the same experimental session*. Our results reveal that measurement error accounts for approximately 50% of the variance of the observed risk-aversion measure, irrespective of the task used to elicit risk-aversion.

The measurement error problem also affects tasks used to elicit other important behavioral measures, such as time preferences or social preferences. For instance, correlation coefficients for individual choices in test/retest measures of time preference are lower than 0.70, as documented by Wölbert and Riedl (2013), Chuang and Schechter (2015), Meier and Sprenger (2015). The existence of a substantial amount of noise is confirmed by Blavatskyy and Maafi (2018) who collected test/retest data within the same experimental session.

The common econometric consequences of including such noisy measures in regressions are (1) a lack of significance of the risk-aversion measure in small samples and (2) biased

coefficients in multivariate regressions. In particular, Pischke (2007) shows that in two-independent variable regressions, one measured with noise and another without noise, if the two measures are positively correlated (but the error is not), then not only the coefficient of the noisy measure is attenuated, but the coefficient of the other variable is spuriously enhanced. For instance, women are often found to be slightly more risk averse than men. Noisy estimates of risk aversion may thus induce gender to become significant while it wouldn't be absent the measurement error (Gillen et al., 2019).

A reasonable empirical strategy to address these measurement-error issues would be to (1) systematically collect test/retest data, which produces unbiased coefficients using ORIV, and (2) balance the cost of increasing the sample size against the risk of finding insignificant coefficients.

Starting at least from Slovic (1964), researchers have realized that elicited measures of risk attitudes are extremely volatile. More than five decades later, within-subject variability across different tasks appears to be a robust phenomenon (see, among others, Deck et al., 2013; Crosetto and Filippin, 2016; Pedroni et al., 2017), and so is variability in test/retest data with the same task over a longer time period (Mata et al., 2018). Why exactly individual choices are so unstable is still a matter of debate. Our results suggest that taming the noise associated with risk-elicitation tasks by using a particular task and/or eliciting additional controls might not be the answer. Researchers should thus anticipate the consequences of considerable measurement error when designing their protocols.

References

Andersen, S., G. W. Harrison, M. I. Lau, and E. Elisabet Rutström (2008). Lost in state space: are preferences stable? *International Economic Review 49*(3), 1091–1112.

Andreoni, J. and W. Harbaugh (2009). Unexpected utility: Experimental tests of five key questions about preferences over risk.

Attanasi, G., N. Georgantzís, V. Rotondi, and D. Vigani (2018). Lottery-and survey-based risk attitudes linked through a multichoice elicitation task. *Theory and Decision 84*(3), 341–372.

Bardsley, N., R. Cubitt, G. Loomes, P. Moffat, C. Starmer, and R. Sugden (2010). *Experimental Economics: Rethinking the Rules*. Princeton University Press.

Beauchamp, J. P., D. Cesarini, and M. Johannesson (2017). The psychometric and empirical properties of measures of risk preferences. *Journal of Risk and Uncertainty 54*(3), 203–237.

Blavatskyy, P. R. and H. Maafi (2018). Estimating representations of time preferences and models of probabilistic intertemporal choice on experimental data. *Journal of Risk and Uncertainty 56*(3), 259–287.

Charness, G., T. Garcia, T. Offerman, and M. C. Villeval (2020). Do measures of risk attitude in the laboratory predict behavior under risk in and outside of the laboratory? *Journal of Risk and Uncertainty 60*(2), 99–123.

Chuang, Y. and L. Schechter (2015). Stability of experimental and survey measures of risk, time, and social preferences: A review and some new results. *Journal of Development Economics 117*, 151–170.

Crosetto, P. and A. Filippin (2013). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty 47*(1), 31–65.

Crosetto, P. and A. Filippin (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics 19*(3), 613–641.

Csermely, T. and A. Rabas (2016). How to reveal people's preferences: Comparing time consistency and predictive power of multiple price list risk elicitation methods. *Journal of Risk and Uncertainty 53*(2-3), 107–136.

Deck, C., J. Lee, J. Reyes, and C. Rosen (2013, 03). A failed attempt to explain within subject variation in risk taking behavior using domain specific risk attitudes. *Journal of Economic Behavior & Organization 87*, 1–24.

Dulleck, U., J. Fooken, and J. Fell (2015). Within-subject intra-and inter-method consistency of two experimental risk attitude elicitation methods. *German Economic Review 16*(1), 104–121.

Engel, C. and O. Kirchkamp (2019). How to deal with inconsistent choices on multiple price lists. *Journal of Economic Behavior & Organization 160*, 138–157.

García-Gallego, A., N. Georgantzís, D. Navarro-Martínez, and G. Sabater-Grande (2011). The stochastic component in choice and regression to the mean. *Theory and Decision 71*(2), 251–267.

Gillen, B., E. Snowberg, and L. Yariv (2019). Experimenting with measurement error: Techniques with applications to the Caltech cohort study. *Journal of Political Economy 127*(4), 1826–1863.

Hey, J. D., A. Morone, and U. Schmidt (2009). Noise and bias in eliciting preferences. *Journal of Risk and Uncertainty 39*(3), 213–235.

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92*(5), 1644–1655.

Jacobson, S. and R. Petrie (2009). Learning from mistakes: What do inconsistent choices over risk tell us? *Journal of Risk and Uncertainty 38*(2), 143–158.

Lönnqvist, J.-E., M. Verkasalo, G. Walkowitz, and P. C. Wichardt (2015). Measuring individual risk attitudes in the lab: Task or ask? An empirical comparison. *Journal of Economic Behavior and Organization 119*, 254–266.

Mata, R., R. Frey, D. Richter, J. Schupp, and R. Hertwig (2018). Risk preference: A view from psychology. *Journal of Economic Perspectives 32*(2), 155–72.

Meier, S. and C. D. Sprenger (2015). Temporal stability of time preferences. *Review of Economics and Statistics 97*(2), 273–286.

O'Donoghue, T. and J. Somerville (2018). Modeling risk aversion in economics. *Journal of Economic Perspectives 32*(2), 91–114.

Pedroni, A., R. Frey, A. Bruhin, G. Dutilh, R. Hertwig, and J. Rieskamp (2017). The risk elicitation puzzle. *Nature Human Behaviour 1*(11), 803–809.

Pischke, S. (2007). Lecture notes on measurement error. *mimeo, London School of Economics, London.*

Sabater-Grande, G. and N. Georgantzis (2002). Accounting for risk aversion in repeated prisoners' dilemma games: An experimental test. *Journal of Economic Behavior & Organization 48*(1), 37–50.

Schildberg-Hörisch, H. (2018). Are risk preferences stable? *Journal of Economic Perspectives 32*(2), 135–54.

Slovic, P. (1964). Assessment of risk taking behavior. *Psychological Bulletin 61*(3), 220.

Wölbert, E. and A. Riedl (2013). Measuring time and risk preferences: Reliability, stability, domain specificity.

Zhou, W. and J. Hey (2018). Context matters. *Experimental Economics 21*(4), 723–756.

Appendix 3.A Simulations for larger samples

In this Appendix we provide estimates for "large" samples: N=300, N=500 and N=1000 (over 10000 simulations).

Table 3.9: Simulations: Simple OLS and IV with N=300 (10 000 simulations)

	β	0.15	0.2	0.25	0.3	0.35
x^*	Mean $\hat{\beta}$	.1511	.2002	.2500	.3013	.3504
	(St Dev)	.0574	.0567	.0572	.0562	.0563
	Sig 0.1	84.34%	97.10%	99.63%	99.99%	100.0%
	Sig 0.05	75.64%	93.88%	99.11%	99.94%	100.0%
	Sig 0.01	53.63%	82.77%	96.34%	99.57%	99.99%
$\lfloor x^* \rfloor$	Mean $\hat{\beta}$	.1403	.1855	.2319	.2791	.3251
	(St Dev)	.0553	.0550	.0553	.0542	.0544
	Sig 0.1	81.89%	95.71%	99.42%	99.97%	100.0%
	Sig 0.05	72.71%	91.98%	98.62%	99.90%	99.99%
	Sig 0.01	49.66%	79.02%	94.60%	99.32%	99.95%
$x^* + \epsilon$	Mean $\hat{\beta}$	.0855	.1130	.1416	.1708	.1981
	(St Dev)	.0434	.0429	.0439	.0429	.0434
	Sig 0.1	63.39%	83.85%	94.41%	99.07%	99.79%
	Sig 0.05	51.45%	75.17%	90.16%	97.77%	99.47%
	Sig 0.01	28.33%	52%	75.54%	91.04%	97.33%
$\lfloor x^* + \epsilon \rfloor$	Mean $\hat{\beta}$	.0818	.1083	.1355	.1637	.1897
	(St Dev)	.0425	.0419	.0429	.0419	.0424
	Sig 0.1	62.00%	82.37%	93.53%	98.72%	99.7%
	Sig 0.05	50.00%	73.10%	89.05%	97.10%	99.25%
	Sig 0.01	26.74%	50.24%	73.25%	89.78%	96.52%
IV	Mean $\hat{\beta}$	.1517	.2013	.2516	.3023	.3519
	(St Dev)	.0786	.0790	.0803	.0801	.0811
	Sig 0.1	62.61%	83.24%	94.26%	98.56%	99.77%
	Sig 0.05	49.81%	73.65%	89.50%	97.00%	99.34%
	Sig 0.01	26.36%	50.41%	73.53%	89.89%	96.69%
ORIV	Mean $\hat{\beta}$	.1518	.2011	.2516	.3030	.3517
	(St Dev)	.0695	.0693	.0704	.0704	.0709
	Sig 0.1	72.14%	90.60%	97.77%	99.75%	99.97%
	Sig 0.05	61.13%	84.05%	95.43%	99.08%	99.85%
	Sig 0.01	37.19%	64.74%	85.98%	96.26%	99.26%

The first column indicates the variable or estimation method used in the univariate OLS regression. The first variable is the true x^*, the second the discretization of the true variable, the third considers the effect of noise, and the fourth combines noise and discretization. For the IV estimations, the discrete noisy measure $\lfloor x_1^* + \epsilon \rfloor$ is instrumented by $\lfloor x_2^* + \epsilon \rfloor$. For the ORIV estimations the stack model uses $\lfloor x_j^* + \epsilon \rfloor$ for $j \in \{1, 2\}$. Columns 2 to 6 indicate the average value, standard deviation and significance of β for 10 000 simulations. For instance the ORIV cell for Sig 0.1 and $\beta = 0.15$ is 72.14%, so that the estimated β using ORIV method is significant in 72.14% of the 10 000 regressions at the 10% level when the true β is 0.15.

Table 3.10: Simulation: Simple OLS and IV with N=500 (10 000 simulations)

	β	0.15	0.2	0.25	0.3	0.35
x^*	Mean $\hat{\beta}$	.1498	.1998	.2503	.2998	.3497
	(St Dev)	.0437	.0430	.0441	.0428	.0436
	Sig 0.1	96.00%	99.85%	100.0%	100.0%	100.0%
	Sig 0.05	92.74%	99.54%	99.99%	100.0%	100.0%
	Sig 0.01	80.18%	97.96%	99.94%	100.0%	100.0%
$\lfloor x^* \rfloor$	Mean $\hat{\beta}$	.1389	.1852	.2318	.2781	.324
	(St Dev)	.0420	.0413	.0423	.0412	.0421
	Sig 0.1	94.85%	99.78%	100.0%	100.0%	100.0%
	Sig 0.05	90.99%	99.41%	100.0%	100.0%	100.0%
	Sig 0.01	76.34%	96.75%	99.87%	100.0%	100.0%
$x^* + \epsilon$	Mean $\hat{\beta}$	.0851	.1132	.1419	.1694	.198
	(St Dev)	.0328	.0327	.0337	.0334	.0338
	Sig 0.1	82.59%	96.49%	99.53%	99.97%	100.0%
	Sig 0.05	73.42%	93.02%	98.94%	99.83%	100.0%
	Sig 0.01	50.30%	79.59%	95.05%	99.19%	99.94%
$\lfloor x^* + \epsilon \rfloor$	Mean $\hat{\beta}$	.0816	.1084	.1358	.1622	.1896
	(St Dev)	.0321	.032	.0329	.0327	.0332
	Sig 0.1	81.19%	95.87%	99.42%	99.92%	100.0%
	Sig 0.05	71.69%	91.8%	98.73%	99.81%	100.0%
	Sig 0.01	48.16%	77.44%	93.88%	98.88%	99.94%
IV	Mean $\hat{\beta}$	.1498	.2	.2511	.3005	.3506
	(St Dev)	.0595	.0604	.0619	.0614	.0632
	Sig 0.1	81.03%	95.56%	99.49%	99.97%	100.0%
	Sig 0.05	71.13%	91.32%	98.71%	99.90%	100.0%
	Sig 0.01	47.76%	77.62%	94.15%	99.08%	99.95%
ORIV	Mean $\hat{\beta}$	.1504	.2002	.2512	.3002	.3506
	(St Dev)	.0522	.0526	.0543	.0541	.0551
	Sig 0.1	89.13%	98.72%	99.92%	100.0%	100.0%
	Sig 0.05	82.36%	96.73%	99.73%	100.0%	100.0%
	Sig 0.01	62.10%	88.91%	98.62%	99.84%	100.0%

The first column indicates the variable or estimation method used in the univariate OLS regression. The first variable is the true x^*, the second the discretization of the true variable, the third considers the effect of noise, and the fourth combines noise and discretization. For the IV estimations, the discrete noisy measure $\lfloor x_1^* + \epsilon \rfloor$ is instrumented by $\lfloor x_2^* + \epsilon \rfloor$. For the ORIV estimations the stack model uses $\lfloor x_j^* + \epsilon \rfloor$ for $j \in \{1, 2\}$. Columns 2 to 6 indicate the average value, standard deviation and significance of β for 10 000 simulations. For instance the ORIV cell for Sig 0.1 and $\beta = 0.15$ is 89.13%, so that the estimated β using ORIV method is significant in 89.13% of the 10 000 regressions at the 10% level when the true β is 0.15.

Table 3.11: Simulation: Simple OLS and IV with N=1000 (10 000 simulations)

	β	0.15	0.2	0.25	0.3	0.35
x^*	Mean $\hat{\beta}$	.1502	.2	.2497	.3003	.3498
	(St Dev)	.0312	.0313	.0307	.0306	.0306
	Sig 0.1	99.93%	100.0%	100.0%	100.0%	100.0%
	Sig 0.05	99.83%	100.0%	100.0%	100.0%	100.0%
	Sig 0.01	98.87%	100.0%	100.0%	100.0%	100.0%
$\lfloor x^* \rfloor$	Mean $\hat{\beta}$	.1392	.1853	.2314	.2784	.3244
	(St Dev)	.0301	.0301	.0298	.0296	.0294
	Sig 0.1	99.92%	100.0%	100.0%	100.0%	100.0%
	Sig 0.05	99.53%	100.0%	100.0%	100.0%	100.0%
	Sig 0.01	97.92%	100.0%	100.0%	100.0%	100.0%
$x^* + \epsilon$	Mean $\hat{\beta}$	.0852	.1132	.1415	.1702	.1979
	(St Dev)	.0234	.0233	.0235	.0236	.0237
	Sig 0.1	97.61%	99.94%	100.0%	100.0%	100.0%
	Sig 0.05	95.34%	99.80%	100.0%	100.0%	100.0%
	Sig 0.01	85.75%	98.84%	99.99%	100.0%	100.0%
$\lfloor x^* + \epsilon \rfloor$	Mean $\hat{\beta}$	.0815	.1083	.1355	.1629	.1893
	(St Dev)	.0229	.0228	.0230	.0231	.0233
	Sig 0.1	97.25%	99.89%	100.0%	100.0%	100.0%
	Sig 0.05	94.49%	99.72%	100.0%	100.0%	100.0%
	Sig 0.01	83.80%	98.50%	99.97%	100.0%	100.0%
IV	Mean $\hat{\beta}$	.1501	.2004	.2499	.3006	.3508
	(St Dev)	.0426	.0431	.0429	.0437	.0443
	Sig 0.1	97.17%	99.93%	100.0%	100.0%	100.0%
	Sig 0.05	94.39%	99.76%	100.0%	100.0%	100.0%
	Sig 0.01	83.81%	98.49%	99.91%	100.0%	100.0%
ORIV	Mean $\hat{\beta}$	.1504	.2002	.2502	.3008	.3504
	(St Dev)	.0374	.0377	.0377	.0383	.0386
	Sig 0.1	99.36%	99.98%	100.0%	100.0%	100.0%
	Sig 0.05	98.37%	99.96%	100.0%	100.0%	100.0%
	Sig 0.01	93.04%	99.81%	100.0%	100.0%	100.0%

The first column indicates the variable or estimation method used in the univariate OLS regression. The first variable is the true x^*, the second the discretization of the true variable, the third considers the effect of noise, and the fourth combines noise and discretization. For the IV estimations, the discrete noisy measure $\lfloor x_1^* + \epsilon \rfloor$ is instrumented by $\lfloor x_2^* + \epsilon \rfloor$. For the ORIV estimations the stack model uses $\lfloor x_j^* + \epsilon \rfloor$ for $j \in \{1, 2\}$. Columns 2 to 6 indicate the average value, standard deviation and significance of β for 10 000 simulations. For instance the ORIV cell for Sig 0.1 and $\beta = 0.15$ is 99.36%s, so that the estimated β using ORIV method is significant in 99.36% of the 10 000 regressions at the 10% level when the true β is 0.15.

Appendix 3.B Additional figures

We simulate data from a probit model to confirm that the problem of measurement error continues to be found with this specification.

The simulations are carried out under the following assumptions:

- $y^* = \alpha + \beta x^* + u \qquad x^* \sim \mathcal{N}(m^{ML}, {\sigma_x^{ML}}^2; [0, 10]) \qquad u \sim \mathcal{N}(0, {\sigma_u}^2) \qquad X^* \perp\!\!\!\perp u$
- $x = \lfloor x^* + \epsilon \rfloor \qquad$ with $\epsilon \sim \mathcal{N}(0, {\sigma_\epsilon}^2) \qquad X^* \perp\!\!\!\perp \epsilon \qquad \epsilon \perp\!\!\!\perp u$
- $\sigma_u^2 = 1 \qquad \alpha = -1.5$
- $y = \mathbb{1}\{y^* > 0\}$

Table 3.12: Simulations: Probit and IV Probit with N=100 (10 000 simulations)

	β	0.15	0.2	0.25	0.3	0.35
x^*	Mean $\hat{\beta}$	.155	.2073	.2589	.3101	.3627
	(St Dev)	.1418	.1359	.1313	.1378	.1444
	Sig 0.1	29.76%	47.71%	64.22%	75.74%	84.39%
	Sig 0.05	20.03%	35.02%	51.47%	64.97%	75.3%
	Sig 0.01	6.33%	14.71%	26.83%	39.47%	50.38%
$\lfloor x^* \rfloor$	Mean $\hat{\beta}$	.143	.1916	.2392	.2863	.3344
	(St Dev)	.1361	.1308	.1263	.1318	.1383
	Sig 0.1	28.77%	45.08%	61.39%	72.60%	81.45%
	Sig 0.05	18.61%	32.48%	48.46%	61.55%	71.89%
	Sig 0.01	5.79%	13.4%	24.26%	35.81%	46.11%
$x^* + \epsilon$	Mean $\hat{\beta}$	.088	.1165	.1457	.1721	.1988
	(St Dev)	.1064	.1006	.0983	.1005	.1049
	Sig 0.1	21.88%	32.38%	44.55%	54.89%	62.58%
	Sig 0.05	13.41%	21.84%	32.01%	41.6%	49.7%
	Sig 0.01	3.85%	7.27%	13.21%	19.59%	24.63%
$\lfloor x^* + \epsilon \rfloor$	Mean $\hat{\beta}$	.0837	.1116	.1396	.1646	.19
	(St Dev)	.1039	.0984	.0959	.0985	.1026
	Sig 0.1	20.96%	31.94%	43.75%	53.11%	60.79%
	Sig 0.05	12.82%	20.95%	31.45%	39.91%	48.32%
	Sig 0.01	3.67%	6.85%	12.15%	18.49%	23.17%
IV	Mean $\hat{\beta}$	.1479	.1994	.2452	.2904	.3346
	(St Dev)	.1814	.1703	.1624	.1625	.1648
	Sig 0.1	25.03%	35.46%	46.1%	56.76%	64.71%
	Sig 0.05	17.18%	25.81%	36.05%	45.93%	55.04%
	Sig 0.01	7.95%	12.97%	20.04%	28.09%	35.94%
ORIV	Mean $\hat{\beta}$	.1517	.2035	.2514	.2963	.3398
	(St Dev)	.1657	.1553	.148	.1492	.152
	Sig 0.1	26.35%	39.22%	53.06%	64.69%	73.03%
	Sig 0.05	17.73%	28.71%	41.19%	52.9%	62.54%
	Sig 0.01	7.47%	13.6%	21.96%	31.57%	40.56%

The first column indicates the variable or estimation method used in the univariate probit regression. The first variable is the true x^*, the second the discretization of the true variable, the third considers the effect of noise, and the fourth combines noise and discretization. For the IV estimations, the discrete noisy measure $\lfloor x_1^* + \epsilon \rfloor$ is instrumented by $\lfloor x_2^* + \epsilon \rfloor$. For the ORIV estimations the stack model uses $\left\lfloor x_j^* + \epsilon \right\rfloor$ for $j \in \{1, 2\}$. Columns 2 to 6 indicate the average value, standard deviation and significance of β for 10 000 simulations. For instance the ORIV cell for Sig 0.1 and $\beta = 0.15$ is 26.35%, so that the estimated β using ORIV method is significant in 26.35% of the 10 000 regressions at the 10% level when the true β is 0.15.

Table 3.13: Simulations: Probit and IV Probit with N=200 (10 000 simulations)

	β	0.15	0.2	0.25	0.3	0.35
x^*	Mean $\hat{\beta}$	.1529	.2033	.2534	.3044	.3552
	(St Dev)	.0965	.0926	.0916	.0943	.0994
	Sig 0.1	49.01%	72.14%	88.18%	95.64%	98.31%
	Sig 0.05	35.93%	60.29%	80.80%	91.86%	96.52%
	Sig 0.01	14.99%	35.63%	59.58%	77.52%	86.81%
$\lfloor x^* \rfloor$	Mean $\hat{\beta}$	.1417	.1879	.2344	.2812	.3275
	(St Dev)	.0927	.089	.0877	.0905	.0952
	Sig 0.1	46.28%	69.27%	86.21%	94.14%	97.56%
	Sig 0.05	33.43%	56.89%	77.69%	89.45%	95.07%
	Sig 0.01	14.34%	31.98%	55.21%	73.40%	83.55%
$x^* + \epsilon$	Mean $\hat{\beta}$	.0862	.1138	.1413	.1679	.1955
	(St Dev)	.0721	.069	.0684	.0688	.0722
	Sig 0.1	32.98%	51.07%	67.89%	79.87%	87.68%
	Sig 0.05	22.21%	38.13%	55.92%	70.45%	79.46%
	Sig 0.01	7.99%	17.64%	31.59%	45.80%	57.02%
$\lfloor x^* + \epsilon \rfloor$	Mean $\hat{\beta}$	.0822	.1087	.1353	.1604	.1869
	(St Dev)	.0702	.0677	.0667	.0675	.0704
	Sig 0.1	31.90%	49.47%	66.58%	78.72%	86.14%
	Sig 0.05	21.49%	36.86%	54.70%	68.14%	77.45%
	Sig 0.01	7.63%	16.84%	29.96%	43.56%	54.85%
IV	Mean $\hat{\beta}$	.1501	.1965	.2418	.2874	.3312
	(St Dev)	.1244	.1187	.1156	.1145	.1151
	Sig 0.1	34.36%	51.53%	67.52%	80.04%	86.96%
	Sig 0.05	24.24%	39.90%	56.88%	70.80%	80.07%
	Sig 0.01	10.97%	21.63%	35.36%	50.65%	61.31%
ORIV	Mean $\hat{\beta}$	.1513	.1987	.2451	.2905	.3350
	(St Dev)	.1119	.1073	.1046	.1039	.1057
	Sig 0.1	39.07%	58.68%	76.72%	87.85%	93.12%
	Sig 0.05	27.88%	46.44%	66.00%	80.61%	88.04%
	Sig 0.01	12.44%	25.73%	43.9%	60.54%	72.42%

The first column indicates the variable or estimation method used in the univariate probit regression. The first variable is the true x^*, the second the discretization of the true variable, the third considers the effect of noise, and the fourth combines noise and discretization. For the IV estimations, the discrete noisy measure $\lfloor x_1^* + \epsilon \rfloor$ is instrumented by $\lfloor x_2^* + \epsilon \rfloor$. For the ORIV estimations the stack model uses $\left\lfloor x_j^* + \epsilon \right\rfloor$ for $j \in \{1, 2\}$. Columns 2 to 6 indicate the average value, standard deviation and significance of β for 10 000 simulations. For instance the ORIV cell for Sig 0.1 and $\beta = 0.15$ is 39.07%, so that the estimated β using ORIV method is significant in 39.07% of the 10 000 regressions at the 10% level when the true β is 0.15.

Appendix 3.C Additional statistics and figures

Correlations within session

Table 3.14: Summary of correlations test/retest

Task	No. subjects	Correlation	Task	No. subjects	Correlation
HL1	175	0.5474	AH7	78	0.4586
HL2	78	0.5505	AH8	78	0.6288
AH1	78	0.5368	AH9	78	0.5842
AH2	78	0.4773	SG1	208	0.6330
AH3	78	0.6198	SG2	208	0.4737
AH4	78	0.4044	SG3	208	0.4643
AH5	78	0.4131	SG4	208	0.4082
AH6	78	0.4698	BRET	61	0.4637

The Normality Assumption in Risk-aversion Measures

The maximum-likelihood estimations in the current paper are based on the underlying assumption that the latent variable of interest is normally distributed. We below display the observed distribution and a normal fit for each sample we use. The normality assumption is easier to evaluate when samples are large and the number of possible choices is limited. For example, the observations using the HL method take fewer than 10 possible values (since very few subjects use the extreme values) and the samples are large. We see that the normality assumption is reasonable. By way of contrast, when samples are small and the number of possible values is large, it is unclear how well the normality assumption holds. For instance, when there are over 50 possible values for the variable of interest and few observations, some values are never observed. The fact that the parametric and non-parametric methods yield similar estimates is reassuring regarding the assumptions made for parametric estimation.

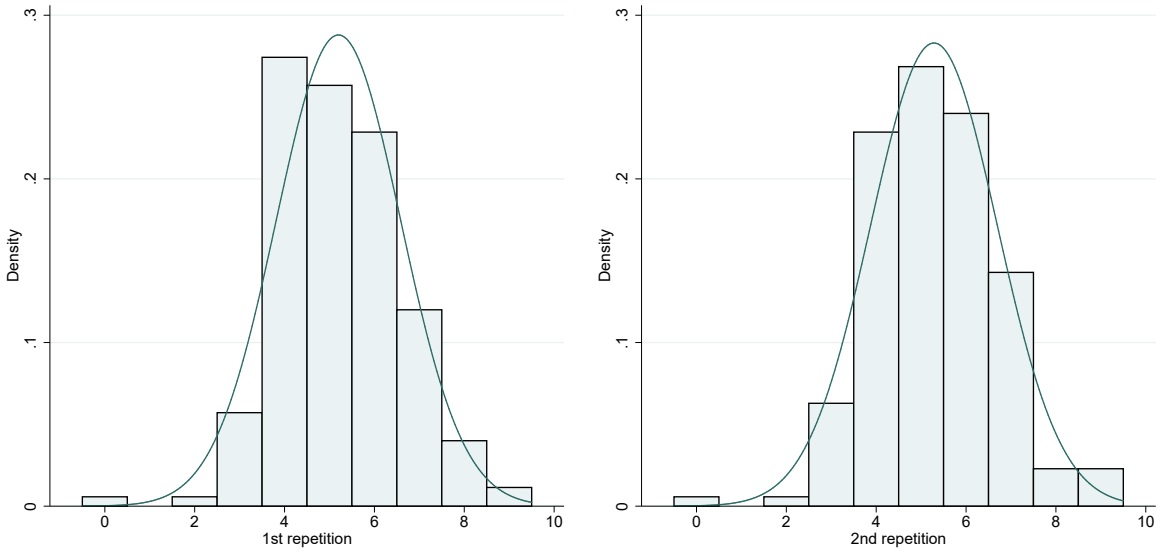


Figure 3.3: The distributions of observed risk attitudes in HL1 and its repetition

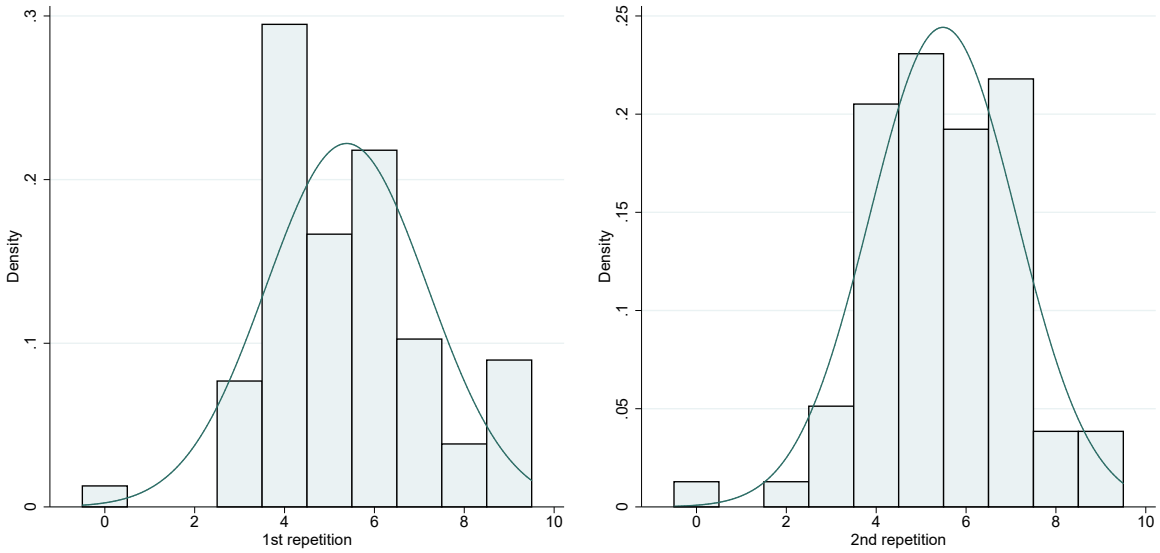


Figure 3.4: The distributions of observed risk attitudes in HL2 and its repetition

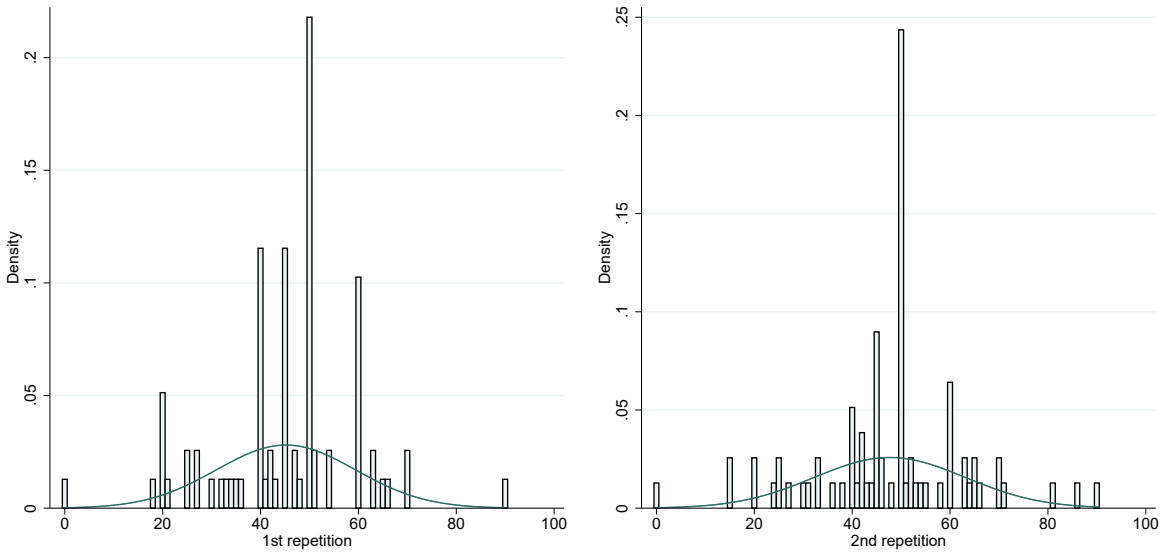


Figure 3.5: The distributions of observed risk attitudes in AH1 and its repetition

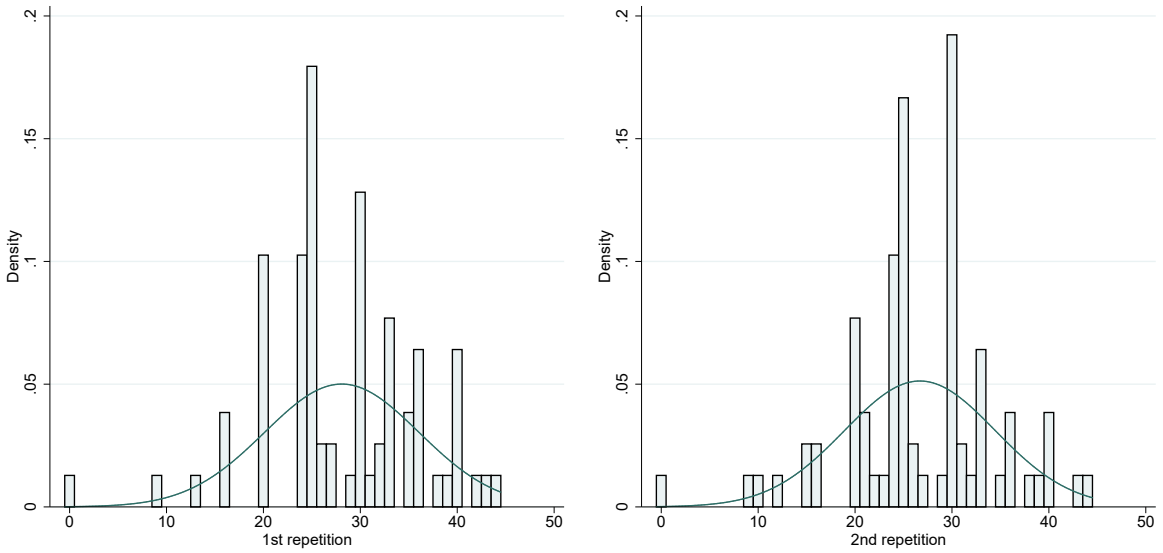


Figure 3.6: The distributions of observed risk attitudes in AH2 and its repetition

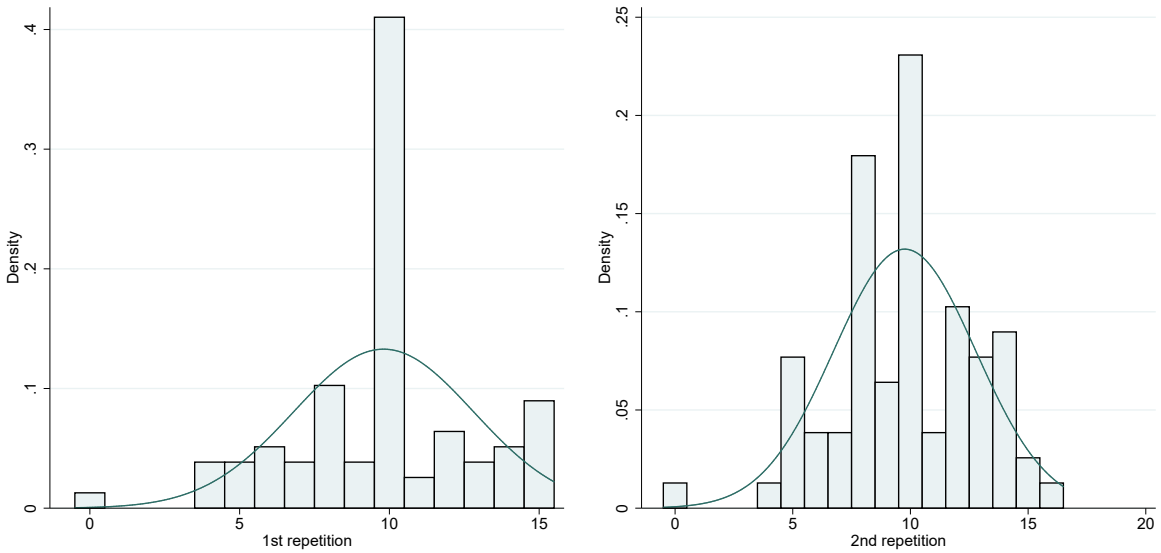


Figure 3.7: The distributions of observed risk attitudes in AH3 and its repetition

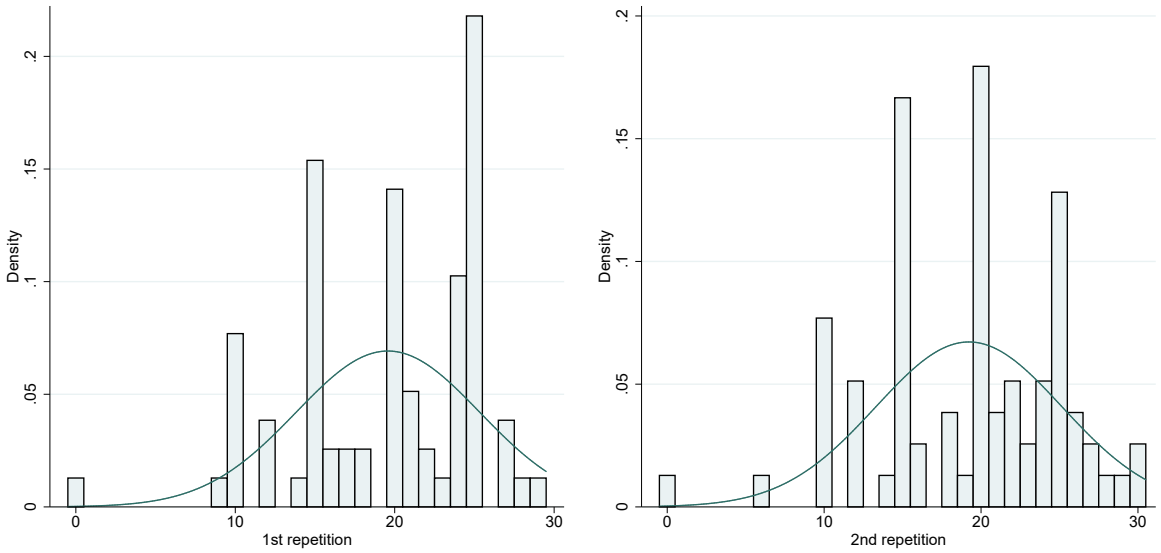


Figure 3.8: The distributions of observed risk attitudes in AH4 and its repetition

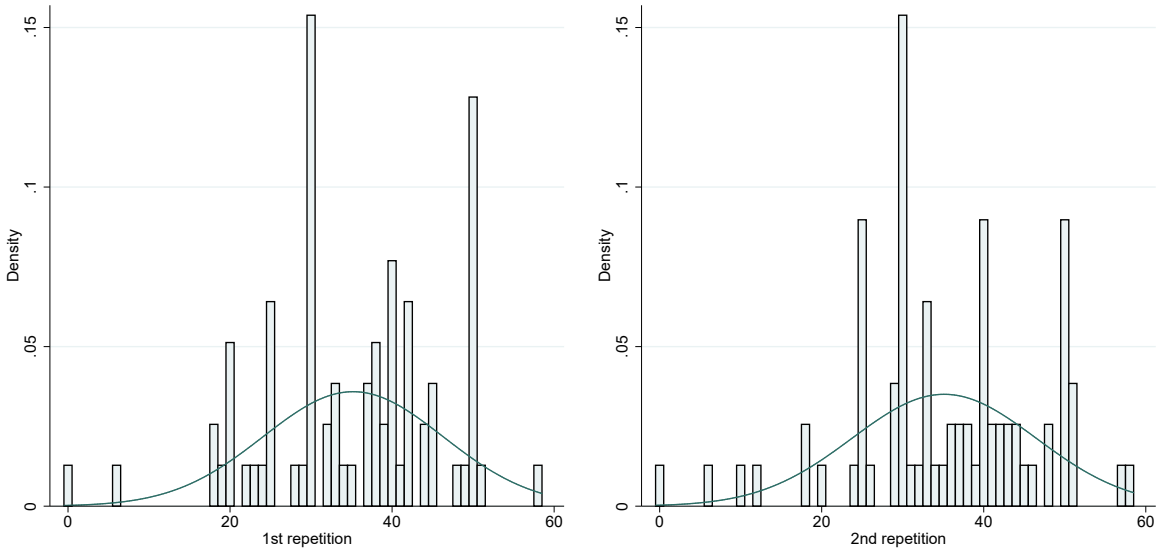


Figure 3.9: The distributions of observed risk attitudes in AH5 and its repetition

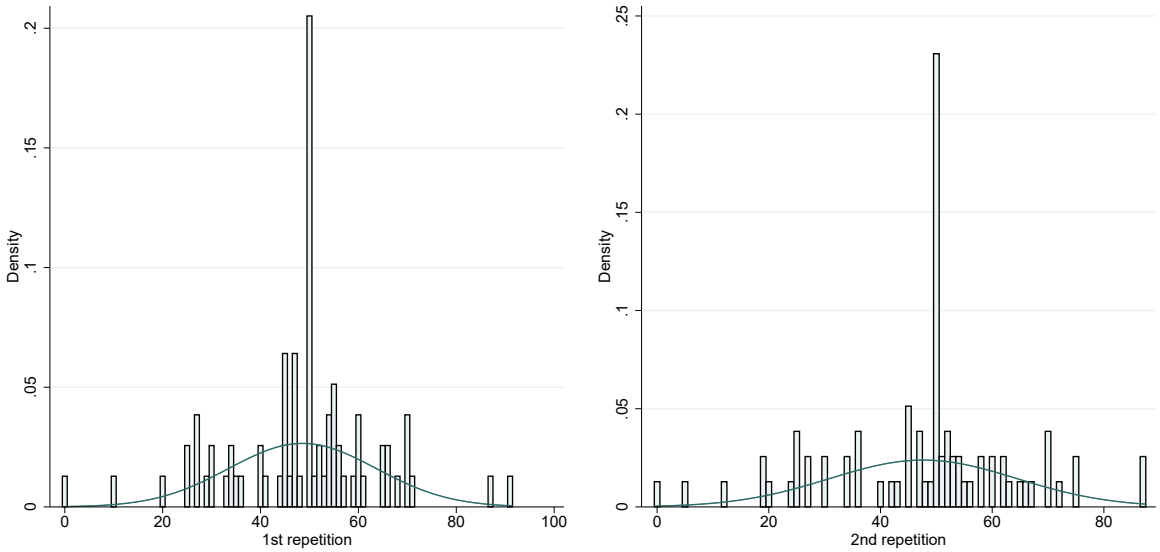


Figure 3.10: The distributions of observed risk attitudes in AH6 and its repetition

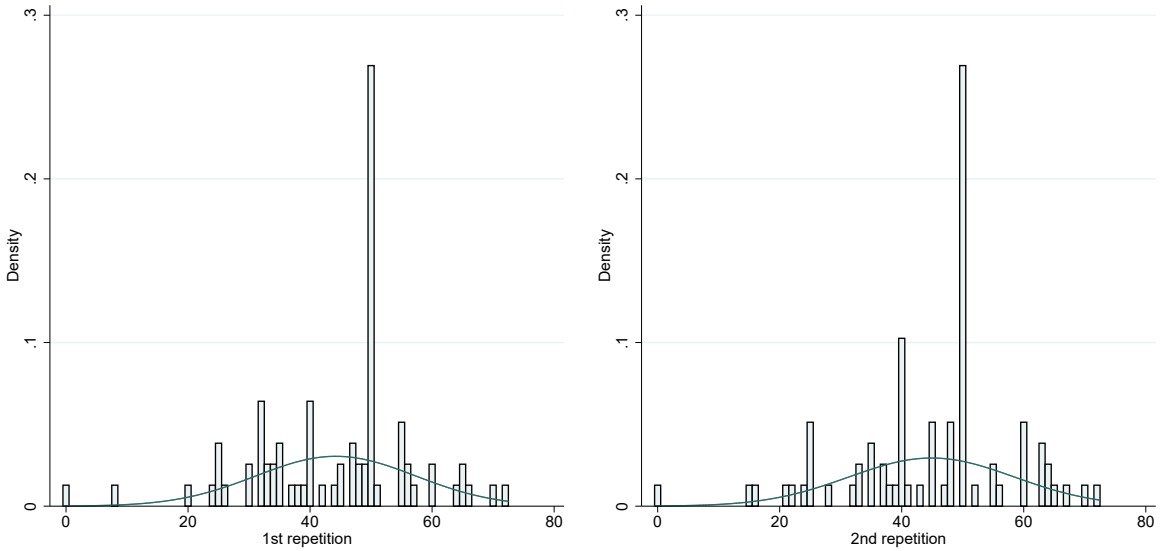


Figure 3.11: The distributions of observed risk attitudes in AH7 and its repetition

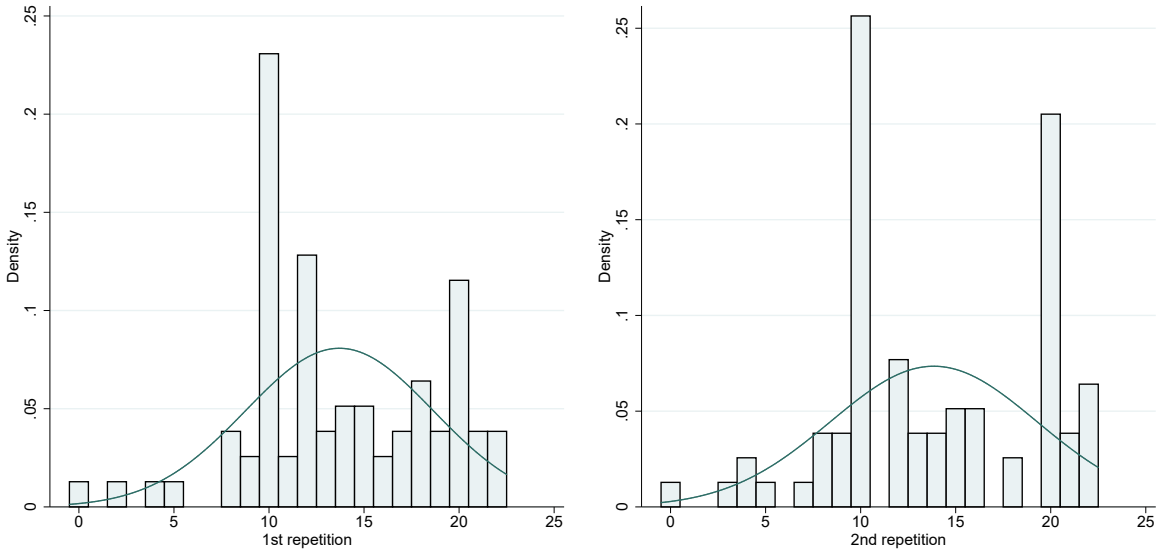


Figure 3.12: The distributions of observed risk attitudes in AH8 and its repetition

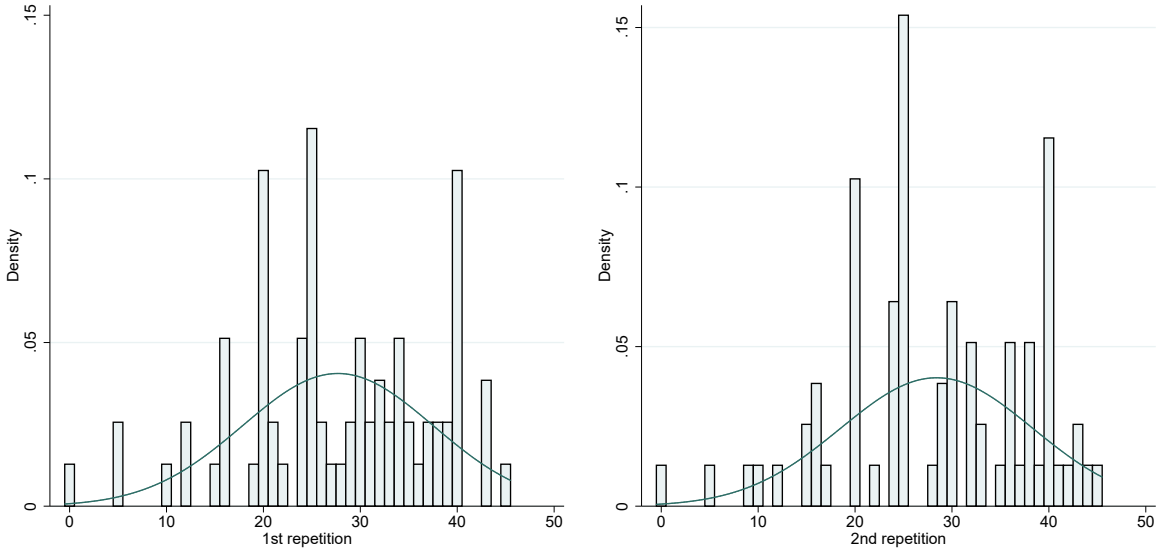


Figure 3.13: The distributions of observed risk attitudes in AH9 and its repetition

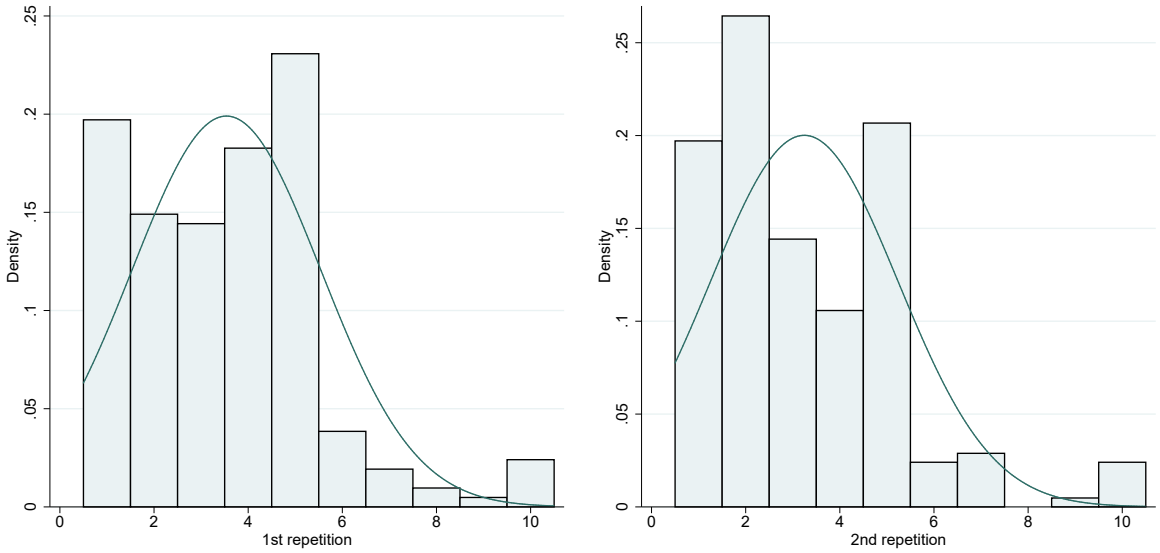


Figure 3.14: The distributions of observed risk attitudes in SG1 and its repetition

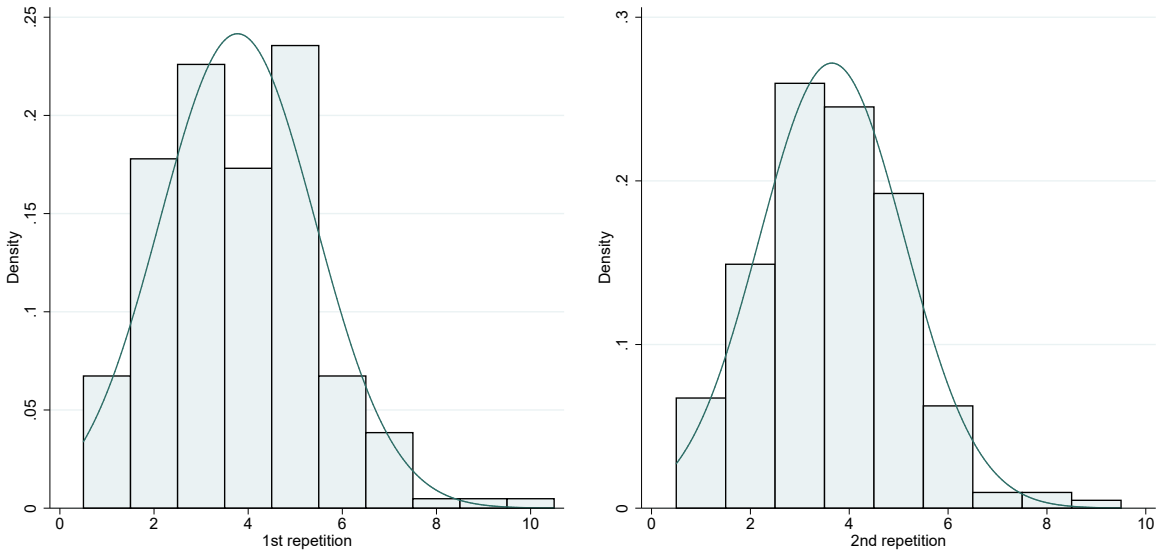


Figure 3.15: The distributions of observed risk attitudes in SG2 and its repetition

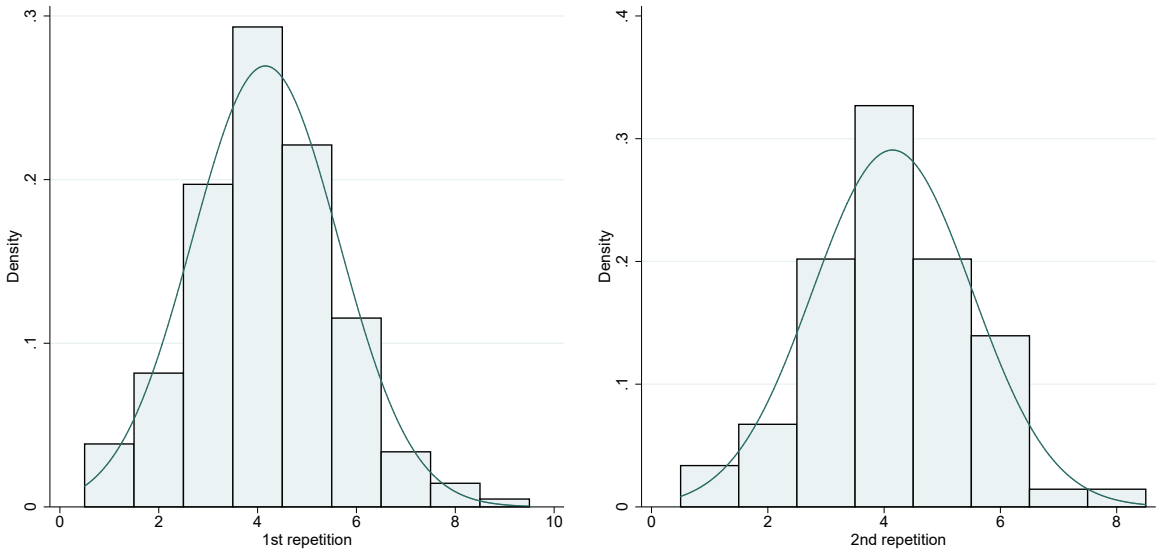


Figure 3.16: The distributions of observed risk attitudes in SG3 and its repetition

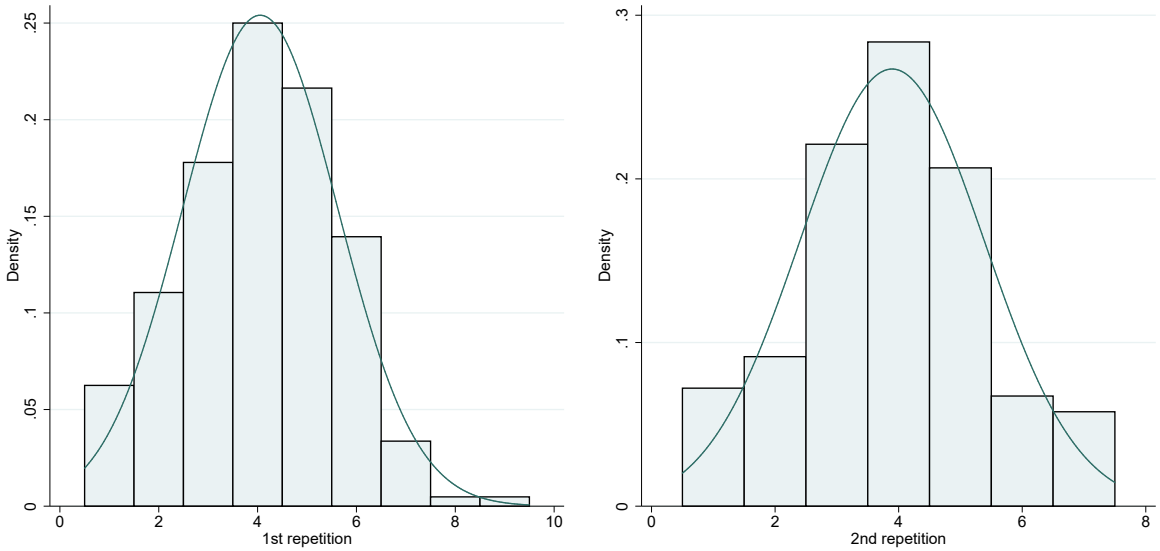


Figure 3.17: The distributions of observed risk attitudes in SG4 and its repetition

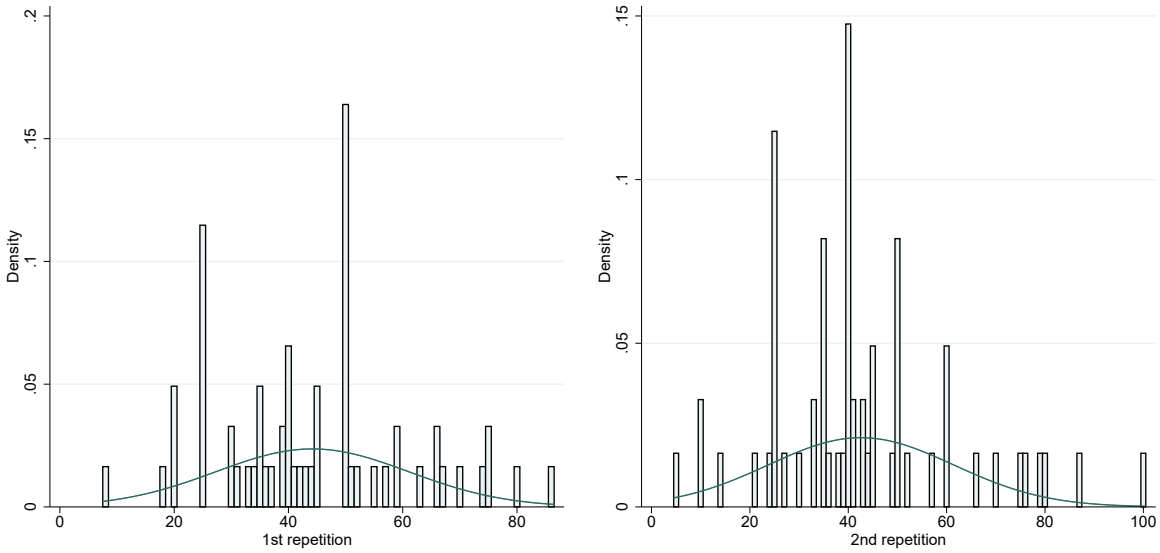


Figure 3.18: The distributions of observed risk attitudes in BRET and its repetition

Chapter 4

Should we trust measures of trust?

Note: This chapter is co-authored with Héloise Clolery, Guillaume Hollard and Inès Picard.[1]

Abstract

Trust is an important economic variable that may however be subject to measurement error, leading to econometric issues such as attenuation bias or spurious correlations. We use a test-retest protocol to assess the measurement error in the two main tasks that are used to elicit trust, namely survey questions and experimental games. We find that trust measures based on the trust game entail substantial measurement error (with up to 15% of noise), while there is virtually no noise in stated trust measures. Given the specificity of our subject pool (students in a top Engineering school) and the short period of time between the test and the retest, we consider this percentage of noise as a lower bound. We also provide a sub-group analysis based on measures of cognitive ability and effort. We find substantial heterogeneity across sub-groups in trust-game behavior, but none for the survey questions. We finally discuss which measure of trust should be used, and the estimation strategies that can be applied to limit the effect of measurement error.

Keywords: Experiments; Test/Retest; Trust; Trust Game; Measurement Error; ORIV.

[1]For this chapter I specifically would like to thank Wael Bousselmi and Lucas Girard as well as participants at the 2021 ESA Global Online Conference and the 2021 ASFEE conference.

1. Introduction

The influential work of Gillen et al. (2019) has drawn considerable attention to the negative consequences of measurement error in experimental work, such as attenuated estimated coefficients in regressions of outcomes on the elicited behavioral characteristics, and biased correlations between these characteristics. The present paper evaluates the extent of measurement error in measures of interpersonal trust. Trust has been shown to have significant economic consequences for a variety of important economic outcomes[1]

To date, little is known about the amount of noise in the standard trust measures that are used in empirical work. In this study, we use a simple test/retest design to gauge the importance of noise in elicited trust measures, and address the following questions. How noisy are they? Are some methods of eliciting trust more prone to measurement error than others? And Can the amount of noise in trust measures be predicted from individual characteristics?

We design a laboratory experiment to elicit a series of trust measures based on the trust game, and also ask survey questions. After brief distracting tasks, the same measures are elicited a second time. This standard test/retest design allows us to gauge the noise included in standard trust measures. We compare the latter using a simple noise ratio.

We find that survey questions are remarkably stable in our test/retest design. However, the measures based on behavior in the trust game are more noisy, with trustworthiness being less noisy than trust measured by the amount sent in the trust game. Compared to similar exercises run on measures of risk attitudes, we find that trust is less noisy than risk. The amount of noise in the measures using the trust game are driven by two individual characteristics: cognitive ability and individual effort (or attention). Subjects with greater cognitive ability produce measures that are stable between tasks. Equally, subjects who put more effort into their work are also less prone to measurement error. One interpretation is that noise can be avoided by subjects who either apply a deliberate strategy in the trust game, or who focus sufficiently on their work to deliver less-noisy answers. In sharp contrast, the measures from survey questions incorporate very little noise, irrespective of the characteristics that appear to lie behind the amount of noise in the game-based measures.

Measurement error attenuates empirical correlations in absolute values. The correlation between different trust measures remains a source of debate.[2] If the various measures

[1] Economic growth (Zak and Knack, 2001; Algan and Cahuc, 2010; Horváth, 2013), the Welfare State (Algan et al., 2016), macroeconomic stability (Sangnier, 2013), international trade and investment (Yu et al., 2015), schooling decisions (Goldin and Katz, 1999), and the performance of large organizations (La Porta et al., 1997).

[2] Previous work has produced ambiguous results. Glaeser et al. (2000), Lazzarini et al. (2003) and Ashraf et al. (2004) find no correlation between stated trust and trust in games, but do find correlations between stated trust and trustworthiness. On the contrary Fehr et al. (2003) and Bellemare and Kröger (2007) find that stated trust is correlated with trust in experimental games but not with trustworthiness. Falk et al. (2016) finds that stated trust can predict trust in games. Thöni et al. (2012) find that trust

of trust reflect a common latent variable then they should be correlated among themselves. However, a number of empirical contributions have found *no* correlation at all between some measures (in particular between survey-based measures and incentivized measures using the trust game). Might measurement error be responsible for these low correlations? Gillen et al. (2019) propose the ORIV estimation method to avoid most of the problems generated by measurement error. We apply this technique to our data, and find only moderate improvements. We do however underline that our subject pool is composed of students in a top Engineering school, i.e. highly-educated subjects who are selected based on their math ability.

We last speculate on recommendations for the trust measure to use in practice. Researchers who wish to limit the amount of noise in their trust measures should definitely use survey questions. Those who attach great importance to incentives should understand that running a trust game in the general population may lead to substantial measurement error.

The remainder of the paper is organized as follows. Section 2 introduces the experimental protocol, Section 3 presents the results and Section 4 concludes.

2. Experimental Design

The purpose of our design is to gauge the amount of noise in elicited measures of trust and risk. We choose a test/retest procedure that consists in having subjects perform the same tasks within a short period of time (about 15 minutes in our experiment). The great advantage of the test/retest design is that it provides very simple non-parametric measures of noise. Participants face the following sequence of tasks (see the detailed descriptions below): (1) Holt and Laury risk measure,[3] (2) Trust Game as the Sender, (3) Trust Game as the Receiver, (4) Holt and Laury measure, (5) Stated Trust, (6) Distracting Tasks, (7) Trust Game as the Sender, (8) Trust Game as the Receiver, (9) Stated Trust. We chose the most widely-used risk and trust measures in the literature. For each incentivized task, payoffs were expressed in coins: 10 coins represent €5.

2.1. Trust Game

We use the standard trust game as proposed by Berg et al. (1995) (also known as the investment game). The trust game is, by far, the most popular incentivized measure

in games is linked to other-regarding behavior while Sapienza et al. (2013) and Banerjee (2018) find on the contrary that it is linked to beliefs about trustworthiness.

[3]Subjects participated in two back-to-back experiments. These were independent (i.e. presented by different experimenters, for an unrelated purpose, were paid independently etc.). The experiment we discuss here is the second one. The first risk-attitude measure was elicited during the other experiment. Having two experiments with the same pool of subjects allows us to compare risk attitudes across experiments.

of trust and trustworthiness (for instance Van Den Akker et al., 2020, find 167 papers that have used this trust game). The Trust game is a two-stage two-player game with a sender and a receiver. In our experiment, both the sender and the receiver have an initial endowment of 10 ECU (i.e. €5). In the first stage, the sender chooses how much of her 10 ECU endowment she wants to send to the receiver (the amounts have to be integers). The amount sent is denoted as X and is the measure of trust. The receiver then obtains X times 3 and chooses Y, the amount she wants to send back to the sender (Y has to be between 0 and $3X$). Y is the measure of trustworthiness. We elicit trustworthiness using the strategy method: subjects are asked to provide a value of Y for the 10 possible values of X corresponding to the integers between 1 and 10. Subjects play each role twice, both times against a random player.

2.2. Stated Trust

A popular way to elicit trust is via a variety of survey questions. We here consider that used, among others, in the World Values Survey (WVS) and the US General Social Survey (GSS). In 2013, Sapienza et al. (2013) counted more than 500 studies using the WVS question: "*Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?*".[4]

The response scale differs across surveys. For instance, the WVS and the GSS propose only two possible answers ("most people can be trusted" and "you can't be too careful"). We here use a 0 to 10 scale, where 0 means "you can't be too careful" and 10 "most people can be trusted". We believe this scale to be more appropriate in detecting variations in test/retest analysis; it has successfully been used in, among others, the European Social Survey. Subjects answered this question twice during the course of the experiment.

2.3. Measure of risk-attitudes: Holt and Laury

Holt and Laury (2002) proposed a measure of risk attitudes via multiple binary choices between lotteries. The Holt-Laury method (HL) is a standard measure of risk-attitudes, and its test/retest properties have been analyzed in a number of contributions (see Perez et al., 2021, for an overview). We compare choices in two Holt-Laury tasks, one from the first experiment and the second from the present experiment. The two tasks are not presented in exactly the same way but involve mathematically-equivalent choices (see the Appendix for screenshots). The amounts involved are roughly equal to those used in the original HL experiment.

[4]The question is slightly differently formulated in the WVS.

2.4. Distracting tasks

Test/retest designs use distracting tasks to avoid making the repetition of identical tasks too salient. We here use a task in which subjects have to count the number of ones in as many matrices as possible over a period of five minutes (as in Figure 4.1). We also included six incentivized questions that are variants of the cognitive reflection test (CRT).[5] We increase the cognitive burden by asking subjects to memorize a seven-digit number (1429587) that they have to report once the distracting tasks are finished. All tasks are incentivized (correctly reporting the number to be memorized is paid 6 coins, each correct count is rewarded by 2 coins with a penalty of -1 coin for incorrect answers, and each correct answer to a CRT question yields 2 coins).

0	0	1	0	0	0	0	0	0	0	0	0
1	0	1	0	0	1	0	1	1	0	0	1
1	1	1	0	1	0	0	0	1	0	1	0
1	1	1	0	0	0	0	0	1	0	0	0
0	1	0	1	1	1	0	1	1	0	1	0
0	0	0	0	1	0	0	0	1	0	1	1
1	0	1	0	0	1	0	1	0	0	1	0
1	0	1	1	0	0	0	0	0	1	1	0

Figure 4.1: Sample of the matrix in the distracting task

3. Results

The experiment took place at CREST Experimental Lab using the O-Tree platform (Chen et al., 2016). We used ORSEE (Greiner, 2004) to recruit 155 students for participation in the experiment. The experiment took place in the Fall of 2021.

3.1. Descriptive statistics

Table 4.1 shows the descriptive statistics for the main variables that appear in our experiment. We carried out Mann-Whitney and Kolmogorov-Smirnov tests to check that the repetitions of each task produce the same distribution: this is the case in our data for all variables apart from the Holt and Laury task. We discuss this in Appendix B. Our

[5]The original CRT was introduced in Frederick (2005): we use variants that correspond to the three first questions in Finucane and Gullion (2010) and questions 4 to 6 from Toplak et al. (2014).

mean values for the amounts sent and the fractions returned in the trust game are similar to those in previous work. We also obtain a similar distribution of transfers, with a large fraction of participants sending either 0 or 10 coins, and a peak at 5.[6] The distributions of our different variables are discussed in Appendix A.

Table 4.1: Descriptive statistics

	No. Obs.	Mean	SD	Min	Max	MW	KS
Amount sent 1	155	4.42	3.63	0	10	.942	1
Amount sent 2	155	4.39	3.68	0	10		
Stated Trust 1	155	4.85	2.55	0	10	.977	1
Stated Trust 2	155	4.85	2.55	0	10		
HL1	155	5.37	1.69	0	9	.003	.049
HL2	155	5.95	1.79	0	10		
Fraction Returned 1	155	.32	.24	0	1	.902	.956
Fraction Returned 2	155	.32	.24	0	1		
CRT	155	4.49	1.61	0	6		
Grid Score	155	13.1	5.53	0	24		
Memorize	155	.93	.26	0	1		

Notes: HL1 (HL2) is the number of safe choices in the first (second) Holt and Laury task. CRT is the number of correct answers to the six CRT-like questions. The Grid Score is twice the number of correct counts of ones minus the number of incorrect counts of ones in the counting task. In the case of negative payoffs, the Grid Score is set to 0. Memorize equals 1 if the subject correctly memorized the 7-digit number.

3.2. Measurement Error

We estimate measurement error in each task using a non-parametric method set out below. We note that rounding issues (i.e. the fact that respondents report integers) is not taken into account. This issue is adressed in Perez et al. (2021), which also proposes non-parametric estimation. We provide similar parametric estimations in Appendix D.[7]

We assume that we observe two noisy measures for each variable x:[8]

$$x_1 = x^* + \epsilon_1 \quad \text{and} \quad x_2 = x^* + \epsilon_2$$

We assume independence between the errors ϵ and the true parameter x^*, and that

[6] See Capra et al. (2008) for a review of experimental results.

[7] Perez et al. (2021) show empirically that parametric and non-parametric methods yield similar estimates of measurement error in four different risk-aversion tasks, using 16 datasets.

[8] Adding a systematic error to the Holt and Laury measures does not change the estimations. See Appendix B and Appendix D for details.

the ϵ are independent and identically-distributed across repetitions.

Then $Var(x_1 - x_2) = Var(\epsilon_1 - \epsilon_2) = 2Var(\epsilon)$, as ϵ_1 and ϵ_2 are assumed to be independent.

$$Var(\epsilon) = \frac{Var(x_1 - x_2)}{2}$$

We can therefore estimate the variance of the noise in measurement using the empirical variances:

$$\widehat{\sigma_\epsilon}^2 = \frac{\widehat{Var}(x_1 - x_2)}{2}$$

The variance of the measure can be estimated by the mean of the empirical variances of x_1 and x_2

$$\widehat{\sigma_m}^2 = \frac{\widehat{Var}(x_1) + \widehat{Var}(x_2)}{2}$$

We are interested in the noise ratio R, defined as the part of the measure's variance that reflects measurement error (noise):

$$R = \frac{\sigma_\epsilon^2}{\sigma_m^2} \qquad \text{or equivalently} \qquad R = 1 - Corr(x_1, x_2)$$

We can estimate this ratio by:

$$\widehat{R} = \frac{\widehat{\sigma_\epsilon}^2}{\widehat{\sigma_m}^2} \qquad \text{or simply by} \qquad \widehat{R} = 1 - \widehat{Corr}(x_1, x_2)$$

Table 4.2: Measurement error

	$\widehat{\sigma_m}^2$	$\widehat{\sigma_\epsilon}^2$	Noise Ratio $\widehat{R}$
Stated Trust	6.5	.32	**4.89%**
Fraction Returned	.057	.0047	**8.24%**
Amount Sent	13.37	2.00	**14.9%**
HL	3.02	1.36	**45.1%**

Notes: For each elicited measure, this table shows the estimated variance $\widehat{\sigma_m}^2$, the estimated variance of the error term $\widehat{\sigma_\epsilon}^2$, and the estimated noise ratio $\widehat{R}$.

We can see in Table 4.2 that the amount of noise varies greatly across measures. The noise ratio in the risk measure reaches a figure of 45%, very much in line with the ratio

found in other datasets (see Perez et al., 2021). By way of contrast, the amounts sent in the trust game appear much less noisy, with a noise ratio of 15%. Trustworthiness, as measured by the average fraction of money that is returned in the trust game, appears to entail only little noise,[9] while at the extreme, stated trust is remarkably stable and (almost) immune to measurement error. These estimations of measurement error (in particular in the trust tasks) should be interpreted as lower bounds of the noise in typical experiments for three reasons. First, the amount of time elapsed between the test-retest measures was quite short. Second, subjects were in a well-equipped experimental laboratory, synonymous with excellent material conditions. Last, our subjects are students in top Engineering schools, and are therefore presumably less prone to noisy behavior.

3.3. Correlations

The ORIV method, proposed by Gillen et al. (2019), allows us to estimate the *true* correlation (i.e. without measurement error), under the assumption of the independence of errors across tasks. To evaluate the impact of measurement error on the correlations, we first calculate the empirical correlations. In our case, with two repeated observations on each variable, we can estimate four different correlations between two variables. Table 4.3 displays the empirical correlations as the mean of all the pairwise correlations. Second, we use the ORIV method to correct for measurement error. These results are also displayed in Table 4.3.

As expected, the ORIV method pushes the correlation up by about 10%. We find that stated trust, trust behavior, and trustworthiness are all significantly and positively correlated. We note that, in our experiment, the correlation figures were already high as compared to those in similar experiments and measurement error quite low, so that these changes are not very significant. As such, stated trust and trust behavior in the trust game seem to reflect a common trait of interpersonal trust.

[9]The fact that we average measures may explain the smaller noise figure for this measure: see Appendix C.

Table 4.3: Empirical and ORIV Correlations

	Stated		Send		Return	
	Emp.	ORIV	Emp.	ORIV	Emp.	ORIV
Stated	1.00	1.00				
Send	.41	.45	1.00	1.00		
	(.062)	(.067)				
Return	.38	.41	.56	.63	1.00	1.00
	(.061)	(.065)	(.053)	(.06)		
Risk	-.03	-.05	-.02	-.03	-.04	-.06
	(.072)	(.102)	(.074)	(.108)	(.07)	(.07)

Notes: The left columns display the mean empirical correlations between pairs of measures, and the right columns show the ORIV correlations between them. Bootstrapped standard errors appear in parentheses

3.4. Sub-group analysis - What drives the noise in experimental measures?

The previous section provides estimates of the *average* amount of noise. We here, in contrast, explore how the amount of noise varies across specific sub-groups. Our sample is obviously limited in size, and covers a particular population (students in Engineering schools) that restricts its external validity. Rather than conducting a systematic review of all of the potential covariates, we focus on two potential drivers of noise identified in previous work: cognitive ability and effort.[10] Cognitive ability is expected to play a more important role in tasks involving strategic uncertainty, and has been found to be correlated with noise (greater cognitive ability leads to less noise). Effort, which can also be considered as a proxy for attention, is also important as poor attention may produce more noise. Furthermore, test/retest designs involving strategic uncertainty are affected by two key variables: (1) The ability to identify a strategy in games, known as strategic ability (e.g. the ability to form beliefs or calculate a best-response), and (2) The level of attention, which makes it more likely to *remember* previous choices. We here measure cognitive ability using the six-question CRT measure, and effort from the score obtained in the real-effort task used as a distractor. We use a median split rule to classify subjects into four categories over two dimensions: cognitive ability and effort/attention. Subjects with a grid score higher (lower) than 12.5 are labeled *High Effort* (*Low Effort*); subjects who have a CRT score higher (lower) than 4.5 are labeled *High CRT* (*Low CRT*). The

[10]For instance, Amador-Hidalgo et al. (2021) find that noise and inconsistent choices in risky tasks are negatively correlated with cognitive ability. In Anderson and Mellor (2009), when subjects are consistent in their answers to survey questions - which they interpret as a measure of their effort or comprehension in the experiment - they also exhibit more stable risk preferences.

following tables, one for each trust measure, list the values for each case. Effort and CRT are relatively uncorrelated, so that the cells contain a similar number of observations. We carry out the same analysis for risk-attitudes to provide a benchmark.

Table 4.4: Estimation of the Noise Ratios in Trust Behavior by Sub-groups

	Low CRT	High CRT	Total
Low Effort	37.4% [19.5%,65.7%] N=30	17.9% [9.8%,31.9%] N=39	24.3% [15.7%,36.6%] N=69
High Effort	16.3% [8.1%,31.3%] N=31	5.9% [3.5%,10%] N=55	8.9% [5.9%,13.3%] N=86
Total	26.4% [16.7%,40.6%] N=61	10.3% [7%,15.2%] N=94	14.9% [11.1%,19.9%] N=155

Notes: The first row in each box corresponds to the error Ratio $\widehat{R}$ estimated for the subjects in that box. The second row is the 95% confidence interval using a Fisher z transformation with the *corrci* command from Cox (2008). The third row lists the number of observations in the box.

Table 4.4 lists the results for trust measured by the sender's behavior in the trust game.[11] The average amount of noise, about 15%, hides large variations across sub-groups. The noise figure, 37%, for the low effort/low CRT group is in particular non-negligible; in sharp contrast, the high effort/high CRT group is almost stable in test/retest. Effort and CRT are thus good predictors of the noise in the trust-game-based measure. For trustworthiness, Table 4.5 reveals a moderate average level of noise. However, noise is almost non-existent for high effort/low CRT (under 2%) but is 15% for the high CRT/low effort sub-group. It is important to note that trustworthiness is elicited using the strategy method, i.e. the reported values are averaged across the potential amounts received. Furthermore, acting in the role of the receiver involves no uncertainty (strategic or otherwise): the outcome of the game is fully determined by the decision made by the receiver. We therefore expect CRT to have less influence than attention in the noise in this measure.

[11] All confidence intervals in this subsection are estimated using a Fisher z transformation. Appendix E provides bootstrapped confidence intervals that are more conservative in our case.

Table 4.5: Estimation of the Noise Ratio in Trustworthiness by sub-groups

	Low CRT	High CRT	Total
Low Effort	5.3% [2.5%,11%] N=30	15.2% [8.2%,27.3%] N=39	10.2% [6.4%,16.1%] N=69
High Effort	1.4% [0.6%,2.8%] N=31	8.3% [4.9%,13.9%] N=55	6.4% [4.2%,9.6%] N=86
Total	3.6% [2.2%,5.9%] N=61	11.3% [7.6%,16.6%] N=94	8.2% [6.1%,11.1%] N=155

Notes: The first row in each box corresponds to the error Ratio $\widehat{R}$ estimated for the subjects in that box. The second row is the 95% confidence interval using a Fisher z transformation with the *corrci* command from Cox (2008). The third row lists the number of observations in the box.

Measures of risk-attitude have been analyzed in detail, and the noise in these measures is typically found to be between 30% and 50%. We here provide corresponding figures to check that our subject pool displays common values, and to use these as a benchmark. Here too, average noise is not very representative of that in the sub-groups. For instance, in the extreme case of low CRT/low effort the risk measures are mostly noise. As such, compared to risk measures trust measures are less prone to noise.

Table 4.6: Estimation of the Noise Ratio in Risk-Attitude measures by sub-groups

	Low CRT	High CRT	Total
Low Effort	92.6% [57.7%,100%] N=30	39.2% [22.5%,63.7%] N=39	61.2% [42.8%,83.3%] N=69
High Effort	44.4% [24.0%,74.9%] N=31	26.1% [16.0%,41.0%] N=55	32.4% [22.3%,45.8%] N=86
Total	68.9% [47.8%,93.6%] N=61	32.1% [22.5%,44.8%] N=94	45.1% [35.0%,57.1%] N=155

Notes: The first row in each box corresponds to the error Ratio $\widehat{R}$ estimated for the subjects in that box. The second row is the 95% confidence interval using a Fisher z transformation with the *corrci* command from Cox (2008). The third row lists the number of observations in the box.

Contrary to the measures discussed above, stated trust is very stable, and varies only little by sub-group. Stated measures of trust are overall stable, as might be expected from

the low average noise figure of 5%.

Table 4.7: Estimation of the Noise Ratio in Stated Trust measures by sub-groups

	Low CRT	High CRT	Total
Low Effort	3.7% [1.7%,7.7%] N=30	6.3% [3.3%,11.8%] N=39	5.1% [3.2%,8.1%] N=69
High Effort	7.3% [3.5%,14.7%] N=31	3.4% [2.0%,5.7%] N=55	4.9% [3.2%,7.4%] N=86
Total	5.4% [3.3%,8.9%] N=61	4.6% [3.0%,6.8%] N=94	4.9% [3.6%,6.7%] N=155

Notes: The first row in each box corresponds to the error Ratio $\widehat{R}$ estimated for the subjects in that box. The second row is the 95% confidence interval using a Fisher z transformation with the *corrci* command from Cox (2008). The third row lists the number of observations in the box.

4. Discussion

Measures of trust are not all created equal regarding the amount of noise they produce. The measure based on the trust game entails more noise than that based on survey questions, but less than that in risk-attitude measures such as the Holt and Laury procedure. Furthermore, stated trust appears stable in test/retest, even for sub-groups that are more prone to noise. Our results thus shed light on the trade-off between incentives and noise for experimenters who wish to measure trust. If incentives are considered as important, the trust game can be used to elicit trust. However, even with highly-educated subjects the amount of noise is not negligible. On the contrary, for an experimenter who is worried about measurement error, surveys of interpersonal trust appear more appropriate.

A back-of-the-envelope calculation suggests that the general population would score lower than our subject pool.[12] The relevant measure of noise in the trust game, according to Table 4.4, would be in the range of 25 to 40%. In particular, "Lab in the field" experiments which target specific populations, sometimes with lower levels of education, and which occur in places with worse material conditions than university experimental laboratories, are likely to produce considerable noise affecting the elicited individual characteristics. Trust seems to be less problematic in this respect than risk-attitudes, but the amounts of noise remain too large to be ignored. Instrumental methods (like the ORIV

[12]The "low CRT" group provided on average 2.8 correct answers out of the six questions. In their meta-analysis Brañas-Garza et al. (2019) find an average score of 1.2 from three questions, which is somewhat lower.

in Gillen et al. (2019)) coupled with larger sample sizes (n>300) are required to solve the issue of measurement error in these tasks.

References

Algan, Y. and P. Cahuc (2010). Inherited trust and growth. *American Economic Review 100*(5), 2060–92.

Algan, Y., P. Cahuc, and M. Sangnier (2016). Trust and the welfare state: the twin peaks curve. *Economic Journal 126*(593), 861–883.

Amador-Hidalgo, L., P. Brañas-Garza, A. M. Espín, T. García-Muñoz, and A. Hernández-Román (2021). Cognitive abilities and risk-taking: Errors, not preferences. *European Economic Review 134*, 103694.

Anderson, L. and J. Mellor (2009, 09). Are risk preferences stable? Comparing an experimental measure with a validated survey-based measure. *Journal of Risk and Uncertainty 39*, 137—160.

Ashraf, N., I. Bohnet, and N. Piankov (2004). Is trust a bad investment? Working Paper, KSG.

Banerjee, R. (2018). On the interpretation of World Values Survey trust question-global expectations vs. local beliefs. *European Journal of Political Economy 55*, 491–510.

Bellemare, C. and S. Kröger (2007). On representative social capital. *European Economic Review 51*(1), 183–202.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity, and social history. *Games and economic behavior 10*(1), 122–142.

Bishara, A. J. and J. B. Hittner (2017). Confidence intervals for correlations when data are not normal. *Behavior Research Methods 49*(1), 294–309.

Brañas-Garza, P., P. Kujal, and B. Lenkei (2019). Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics 82*, 101455.

Capra, C. M., K. Lanier, and S. Meer (2008). Attitudinal and behavioral measures of trust: A new comparison. *Department of Economics, Emory University, Mimeo.*

Chen, D. L., M. Schonger, and C. Wickens (2016). Otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88–97.

Cox, N. J. (2008). Speaking Stata: Correlation with confidence, or Fisher's z revisited. *The Stata Journal 8*(3), 413–439.

Falk, A., A. Becker, T. J. Dohmen, D. Huffman, and U. Sunde (2016). The preference survey module: A validated instrument for measuring risk, time, and social preferences. Technical report, IZA Discussion Paper.

Fehr, E., U. Fischbacher, J. Schupp, B. Rosenbladt, and G. Wagner (2003). A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative surveys. *Schmoller's Jahrbuch 122*, 519–542.

Finucane, M. L. and C. M. Gullion (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychology and Aging 25*(2), 271.

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

Gillen, B., E. Snowberg, and L. Yariv (2019). Experimenting with measurement error: Techniques with applications to the Caltech cohort study. *Journal of Political Economy 127*(4), 1826–1863.

Glaeser, E. L., D. I. Laibson, J. A. Scheinkman, and C. L. Soutter (2000). Measuring trust. *Quarterly Journal of Economics 115*(3), 811–846.

Goldin, C. and L. F. Katz (1999). Human capital and social capital: The rise of secondary schooling in America, 1910-1940. *The Journal of Interdisciplinary History 29*(4), 683–723.

Greiner, B. (2004). An online recruitment system for economic experiments. Technical report, MPRA.

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92*(5), 1644–1655.

Horváth, R. (2013). Does trust promote growth? *Journal of Comparative Economics 41*(3), 777–788.

La Porta, R., F. Lopez-de Silanes, A. Shleifer, and R. W. Vishny (1997). Trust in large organizations. *American Economic Review 87*(2), 333.

Lazzarini, S., R. Madalozzo, R. Artes, and J. Siqueira (2003). Measuring trust: An experiment in Brazil. Technical report, IBMEC.

Perez, F., G. Hollard, and R. Vranceanu (2021). How serious is the measurement-error problem in risk-aversion tasks? *Journal of Risk and Uncertainty*, 319–342.

Sangnier, M. (2013). Does trust favor macroeconomic stability? *Journal of Comparative Economics 41*(3), 653–668.

Sapienza, P., A. Toldra-Simats, and L. Zingales (2013). Understanding trust. *Economic Journal 123*(573), 1313–1332.

Thöni, C., J.-R. Tyran, and E. Wengström (2012). Microfoundations of social capital. *Journal of Public Economics 96*(7-8), 635–643.

Toplak, M. E., R. F. West, and K. E. Stanovich (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning 20*(2), 147–168.

Van Den Akker, O. R., M. A. Van Assen, M. Van Vugt, and J. M. Wicherts (2020). Sex differences in trust and trustworthiness: A meta-analysis of the trust game and the gift-exchange game. *Journal of Economic Psychology 81*, 102329.

Yu, S., S. Beugelsdijk, and J. de Haan (2015). Trade, trust and the rule of law. *European Journal of Political Economy 37*, 102–115.

Zak, P. J. and S. Knack (2001). Trust and growth. *Economic Journal 111*(470), 295–321.

Appendix 4.A Distributions of the main measures

This Appendix plots the distributions of our observations in the repeated measures for each of our main variables. Incentivized trust (Figure 4.3) follows a U-shaped distribution, with many people sending either none or all of their coins to the receiver, and few people sending amounts inbetween.

On the contrary, when answering the survey questions, our subjects are less concentrated on the extremes and appear more in the center of the distribution, producing a near normal distribution (Figure 4.2).

Concerning trustworthiness, the amounts sent back when participants play the role of the receiver are clustered around 0. There is also a peak at around one half, i.e. a fair redistribution of the gains (Figure 4.4).

Last, the Holt and Laury task produces a near normal distribution, as found in previous work (Figure 4.5).

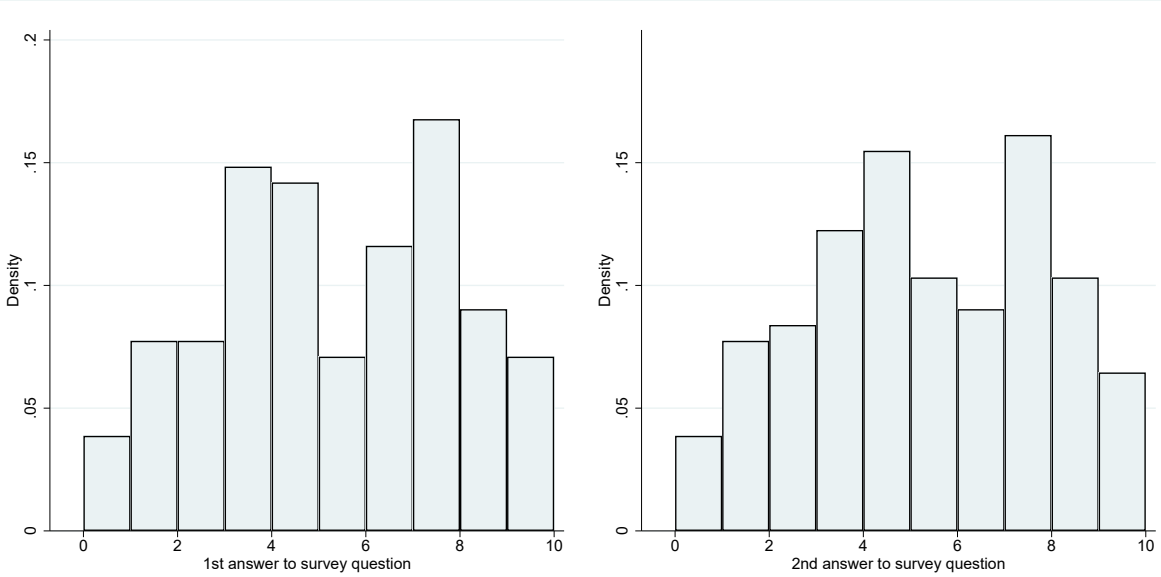


Figure 4.2: The Distribution of Stated Trust in the Two Repetitions

Note: The histogram on the left (right) represents the density of the answers to the survey question the first (second) time it was asked.

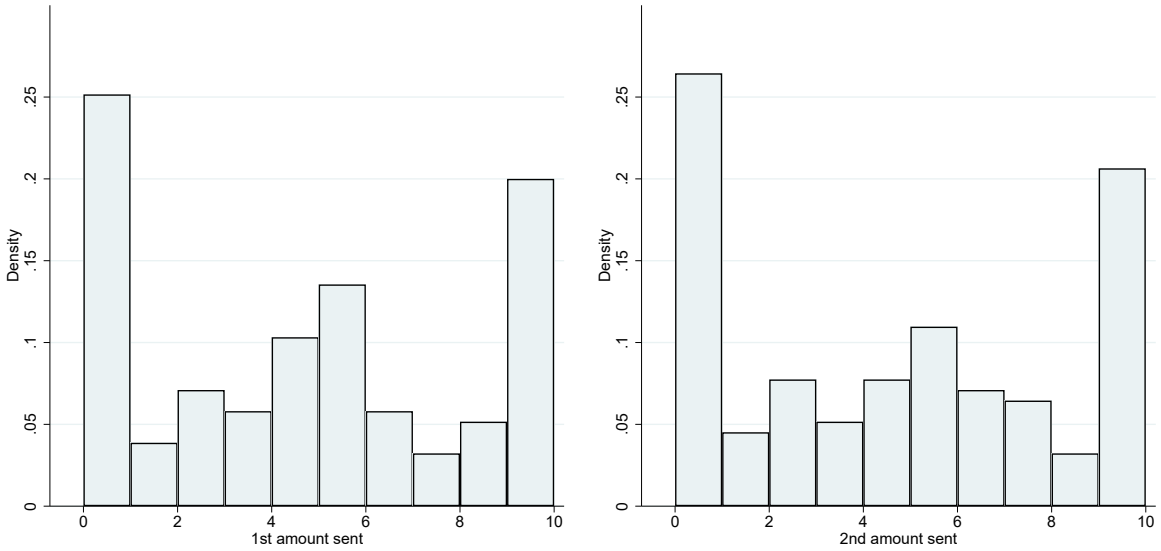


Figure 4.3: The Distribution of the Amount Sent in the Two Repetitions

Note: The histogram on the left (right) represents the density of the amounts sent in the trust game the first (second) time it was played.

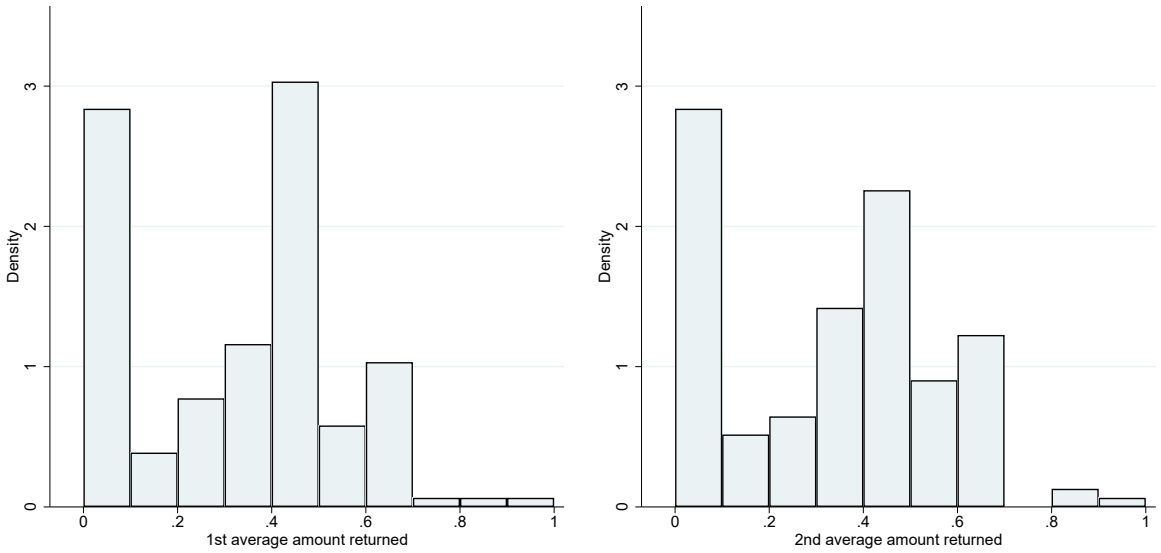


Figure 4.4: The Distribution of the Average Amount Returned in the Two Repetitions

Note: The histogram on the left (right) represents the density of the average shares sent back in the trust game the first (second) time it was played.

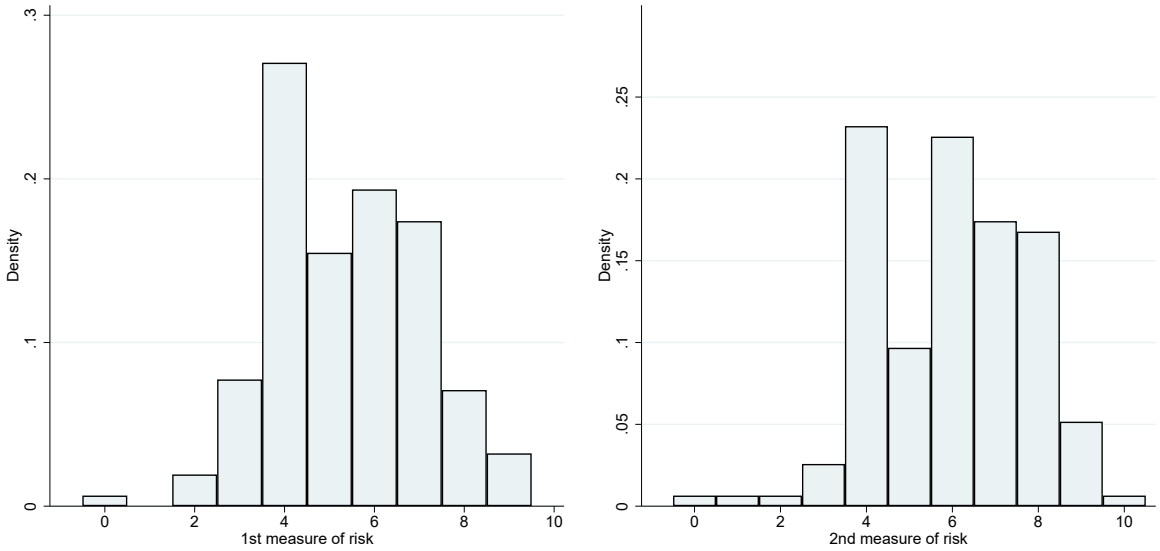


Figure 4.5: The Distribution of Safe Choices in the Two Repetitions of the HL task

Note: The histogram on the left (right) represents the density of the number of safe choices in the first (second) HL task.

Appendix 4.B Repeated observations

This Appendix plots the individuals' repeated observations for our main variables. While the observations are close to the 45 degree line for stated trust (Figure 4.6) and trustworthiness (Figure 4.8), we observe much more noise for incentivized trust (Figure 4.7) and risk (Figure 4.9).

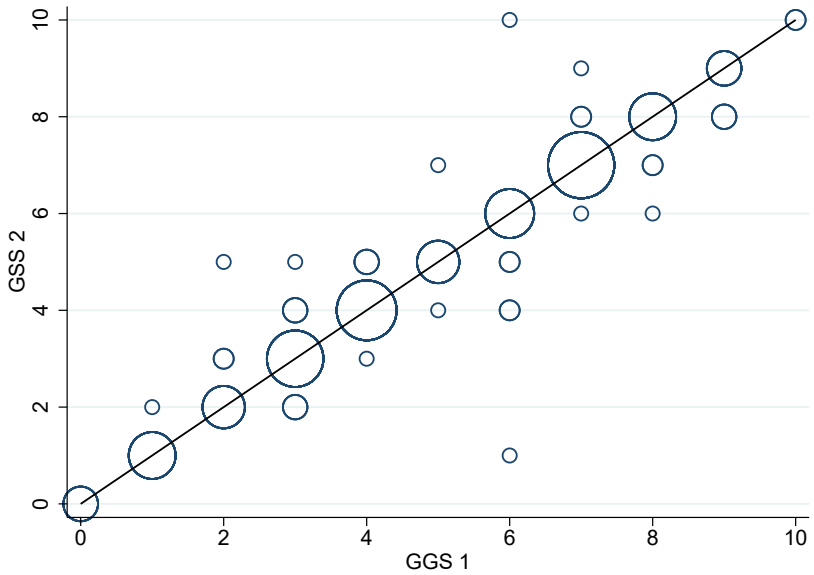


Figure 4.6: Repetition of Stated Trust

Note: This plot represents the repeated answers to the survey question

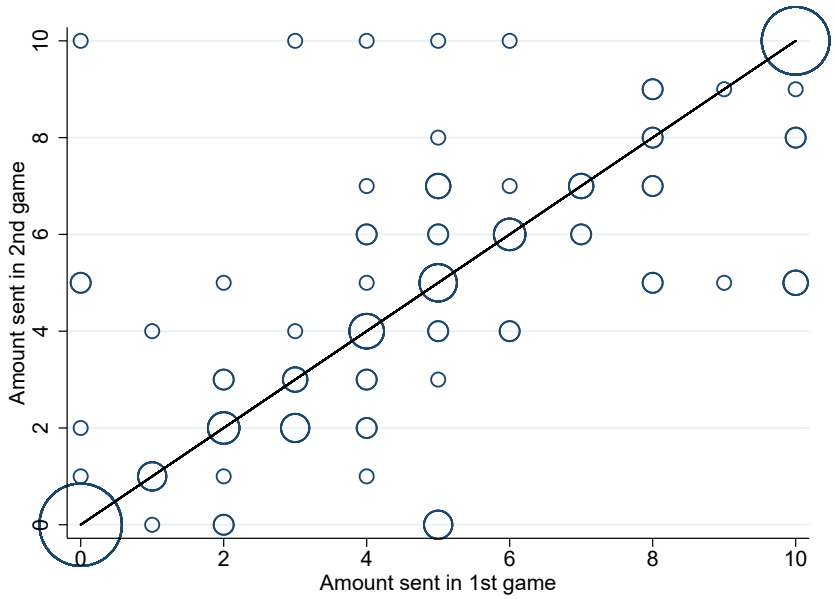


Figure 4.7: Repetition of the Amount Sent

Note: This plot represents the repeated amounts sent in the trust game

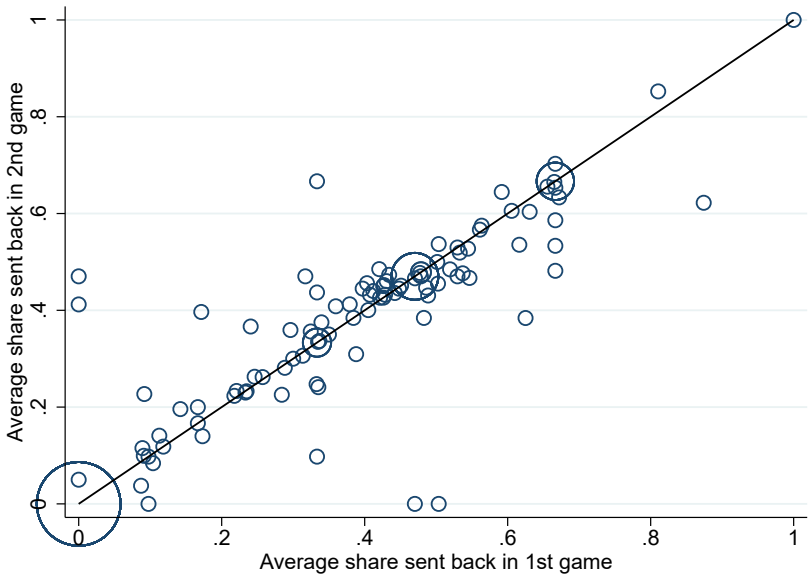


Figure 4.8: Repetition of Trustworthiness

Note: This plot represents the repeated average shares sent back to the sender in the trust game.

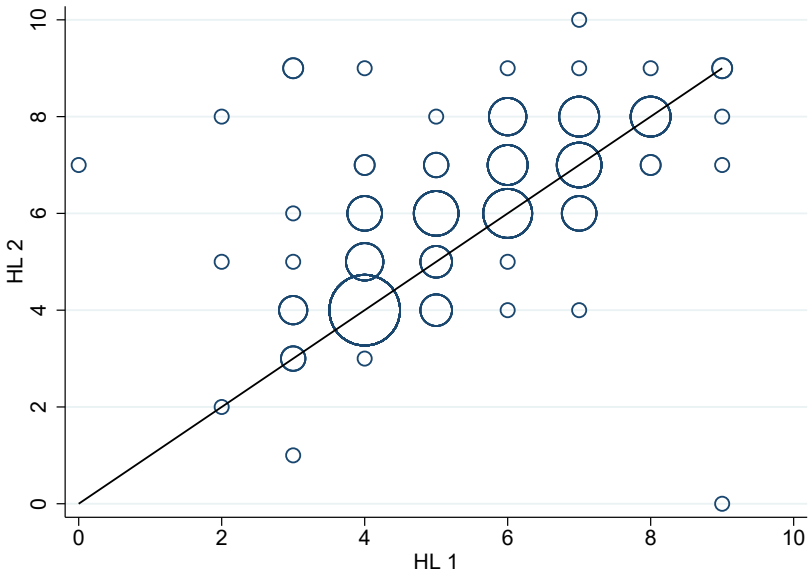


Figure 4.9: Repetition of the HL Task

Note: This plot represents the repeated number of safe choices in the HL task

We use the risk-elicitation task from a previous experiment to estimate measurement error. This previous experiment took place just before the experiment that we analyze in this paper, such that there is about twenty minutes between the two risk-elicitation tasks. The incentives were the same in the two tasks. However, the units in question were

different, with a different conversion rate between ECU and Euros. This may have nudged subjects to choose safer options in the second experiment. Nonetheless, our study focuses on the measurement error given by the random error and not the systematic error that could affect the first or second experiment. In other words, we can model what happens in both risk-elicitation tasks in the following way:

$$x_1 = \lfloor x^* + \epsilon_1 \rfloor \qquad \text{and} \qquad x_2 = \lfloor c + x^* + \epsilon_2 \rfloor$$

where c is the systematic error (a constant) that affects the second risk-elicitation task in comparison to the first one, and ϵ_1 and ϵ_2 are the two random errors. As we focus on random error, the fact that we have this systematic error is not a problem. Furthermore, parametric estimations that take this systematic error into account yield similar results to the non-parametric estimation we discuss in the paper (See Appendix C)

Appendix 4.C Trustworthiness

As discussed in the paper, the trustworthiness variable is the average proportion sent back by the receiver in the trust game. The ratio of measurement error observed in Table 4.2 is quite small, which can reflect that we consider the average of 10 figures. Constructing ten variables (Ret_x) corresponding to the trustworthiness for each choice in the strategy method (how many ECU the receiver sends back when the sender sends x ECU for $x \in [1; 10]$). Table 4.8 show the descriptive statistics for the 10 return variables.

Table 4.8: Amount returned in the ten cases in the strategy method

	First repetition	Second repetition
Ret_1	.75	.8
Ret_2	1.55	1.59
Ret_3	2.23	2.27
Ret_4	3.06	3.01
Ret_5	3.73	3.79
Ret_6	4.44	4.45
Ret_7	5.16	5.21
Ret_8	5.99	6.02
Ret_9	6.72	6.8
Ret_{10}	7.71	7.84

Table 4.9: Measurement error estimated for the different return variables

	$\widehat{\sigma_m}^2$	$\widehat{\sigma_\epsilon}^2$	$\widehat{R}$
Ret_1	.60	.10	.16
Ret_2	2.46	.29	.12
Ret_3	5.08	.46	.09
Ret_4	9.20	.93	.1
Ret_5	14.15	1.27	.09
Ret_6	19.75	1.76	.09
Ret_7	26.89	2.43	.09
Ret_8	36.08	3.36	.09
Ret_9	45.7	5.05	.11
Ret_{10}	60.45	7.42	.12

Appendix 4.D Parametric Estimation of measurement error

Using a similar method to that in Perez et al. (2021), we here provide the estimated variances and noise ratios using a parametric approach. We calculate these for stated trust, the amount sent in the Trust Game, and the Holt and Laury measure (allowing for a systematic error in HL). As noted above, as the return variable is the average of 10 figures, we have decided not to provide any parametric estimation of the measurement errors in this task. Parametric estimations lead to similar result, with a very slightly lower noise ratio being estimated in these three tasks.

Table 4.10: Measurement Error Parametric

	$\widehat{\sigma_m}^2$	$\widehat{\sigma_\epsilon}^2$	Noise Ratio $\widehat{R}$
Stated Trust	6.11	0.229	**3.60%**
Amount Sent	13.76	2.11	**13.3%**
HL	3.02	1.32	**43.9%**

Appendix 4.E Sub-group Analysis with Bootstrapped Confidence Intervals

There is an ongoing debate about the method that should be used to calculate confidence intervals for empirical correlations (see the recent paper by Bishara and Hittner,

2017). The Fisher z transformation to obtain confidence intervals produces not very conservative figures, in particular when the variables are not Gaussian. Here our variables are bounded, so the confidence intervals may not be totally inappropriate. We nonetheless provide the sub-group analysis with bootstrapped confidence intervals that are less conservative (and thus larger, especially with small n).

Table 4.11: Estimation of the Noise Ratio in Trust Behavior by Sub-groups

	Low CRT	High CRT	Total
Low ScoreGrid	37.4% [3.1%,71.7%] N=30	17.9% [6.8%,29.1%] N=39	24.3% [9.8%,38.8%] N=69
High ScoreGrid	16.3% [2.8%,29.8%] N=31	5.9% [0.1%,11.8%] N=55	8.9% [3%,14.8%] N=86
Total	26.4% [8.1%,44.8%] N=61	10.3% [4.8%,15.9%] N=94	14.9% [8.2%,21.7%] N=155

Notes: The first row in each box corresponds to the error Ratio $\widehat{R}$ estimated on the subjects in that box. The second row is the 95% Bootstrap confidence interval, and the third row the number of observations.

Table 4.12: Estimation of the Noise Ratio in Stated Trust by Sub-groups

	Low CRT	High CRT	Total
Low ScoreGrid	3.7% [0.3%,7.0%] N=30	6.3% [0.0%,13.3%] N=39	5.1% [1.1%,9.1%] N=69
High ScoreGrid	7.3% [0.0%,19.0%] N=31	3.4% [0.0%,6.8%] N=55	4.9% [0.0%,9.9%] N=86
Total	5.4% [0.0%,11.2%] N=61	4.6% [1.1%,8.0%] N=94	4.9% [1.7%,8.1%] N=155

Notes: The first row in each box corresponds to the error Ratio $\widehat{R}$ estimated on the subjects in that box. The second row is the 95% Bootstrap confidence interval, and the third row the number of observations.

Table 4.13: Estimation of the Noise Ratio in the Risk-Attitude Measure by Sub-groups

	Low CRT	High CRT	Total
Low ScoreGrid	92.6% [26.4%,100%] N=30	39.2% [13.3%,65%] N=39	61.2% [25.2%,97.2%] N=69
High ScoreGrid	44.4% [15.0%,73.9%] N=31	26.1% [6.9%,45.3%] N=55	32.4% [16.2%,48.5%] N=86
Total	68.9% [29.0%,100%] N=61	32.1% [16.3%,48.0%] N=94	45.1% [25.8%,64.3%] N=155

Notes: The first row in each box corresponds to the error Ratio $\widehat{R}$ estimated on the subjects in that box. The second row is the 95% Bootstrap confidence interval, and the third row the number of observations.

Table 4.14: Estimation of the Noise Ratio in Trustworthiness by Sub-groups

	Low CRT	High CRT	Total
Low ScoreGrid	5.3% [0.8%,9.8%] N=30	15.2% [1.4%,29%] N=39	10.2% [2.7%,17.7%] N=69
High ScoreGrid	1.4% [0.2%,2.5%] N=31	8.3% [0.0%,18.0%] N=55	6.4% [0.0%,13.3%] N=86
Total	3.6% [1.1%,6.1%] N=61	11.3% [2.8%,19.8%] N=94	8.2% [3.0%,13.5%] N=155

Notes: The first row in each box corresponds to the error Ratio $\widehat{R}$ estimated on the subjects in that box. The second row is the 95% Bootstrap confidence interval, and the third row the number of observations.

Appendix 4.F Experiment Screenshots

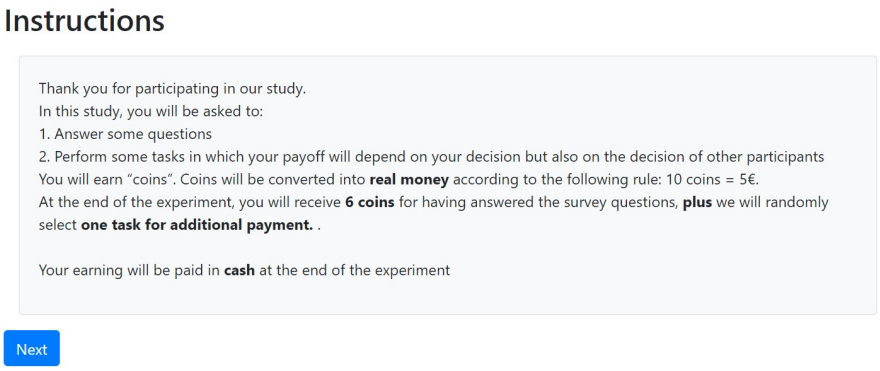


Figure 4.10: Instructions for All Experiments

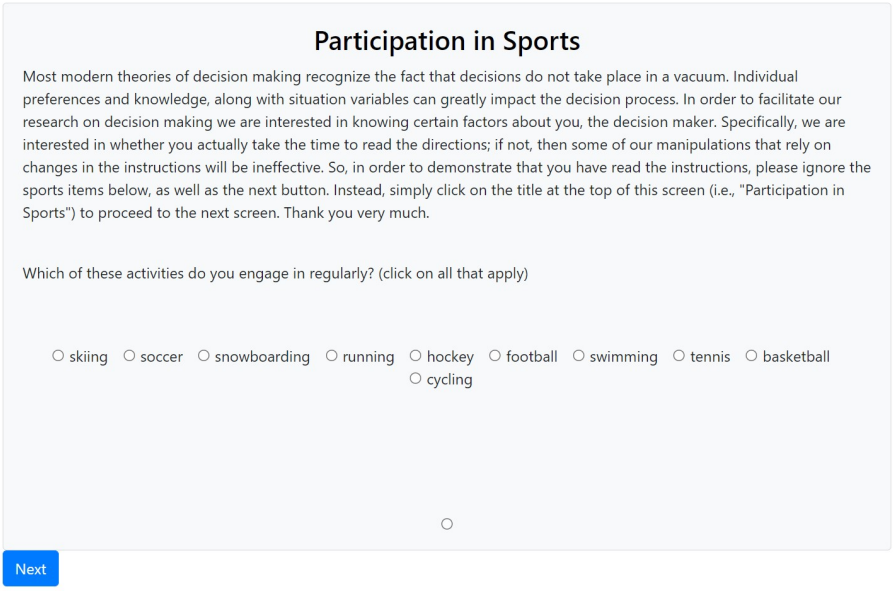


Figure 4.11: Question to Focus Participants' Attention

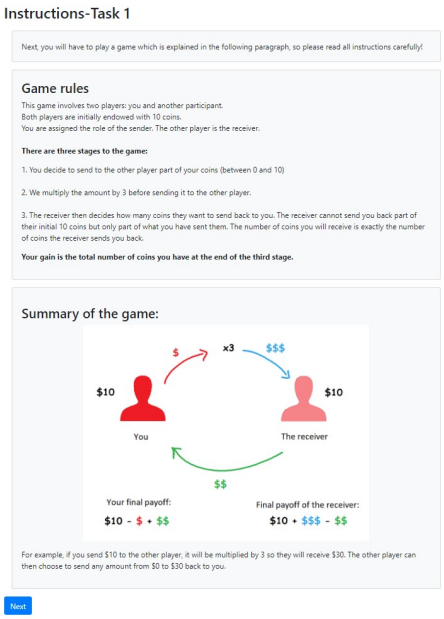


Figure 4.12: Trust Game Instructions

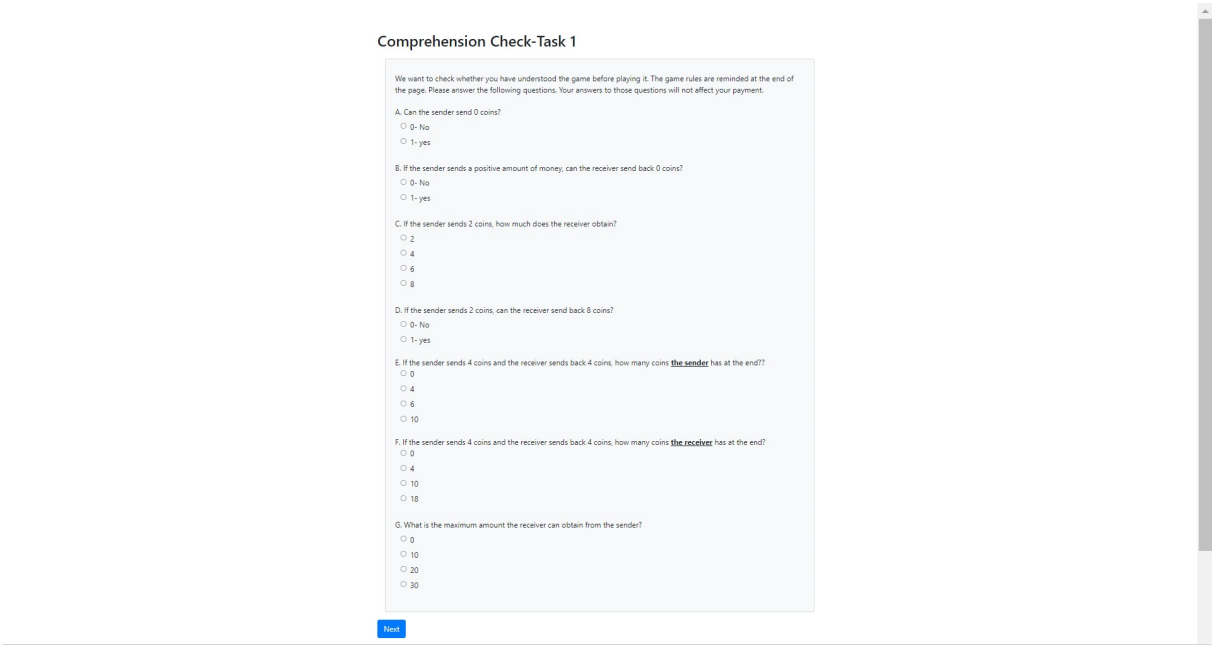


Figure 4.13: Trust Game Comprehension Check

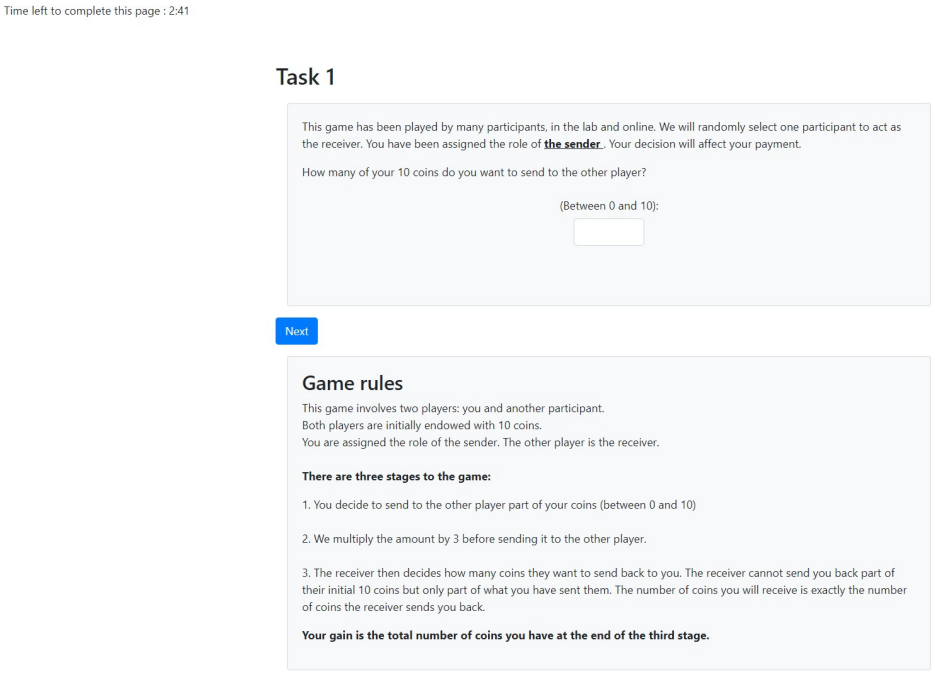


Figure 4.14: Answer to Trust Game as Sender

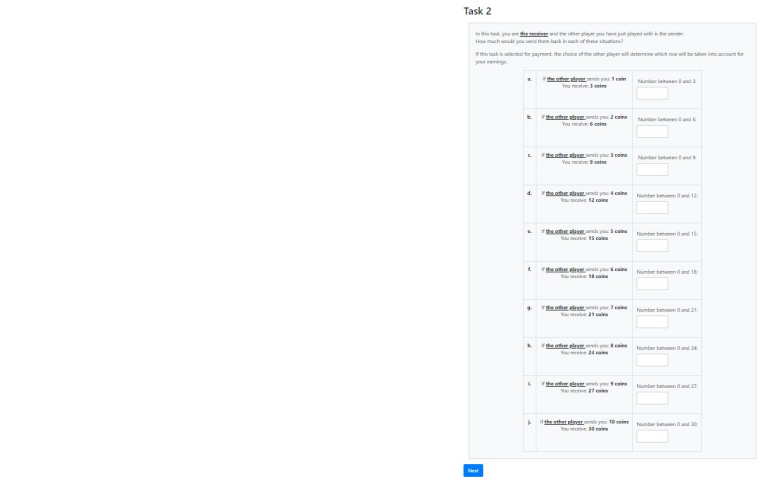


Figure 4.15: Answer to Trust Game as Receiver

Figure 4.16: Instructions for the Risk-Preference Elicitation Task

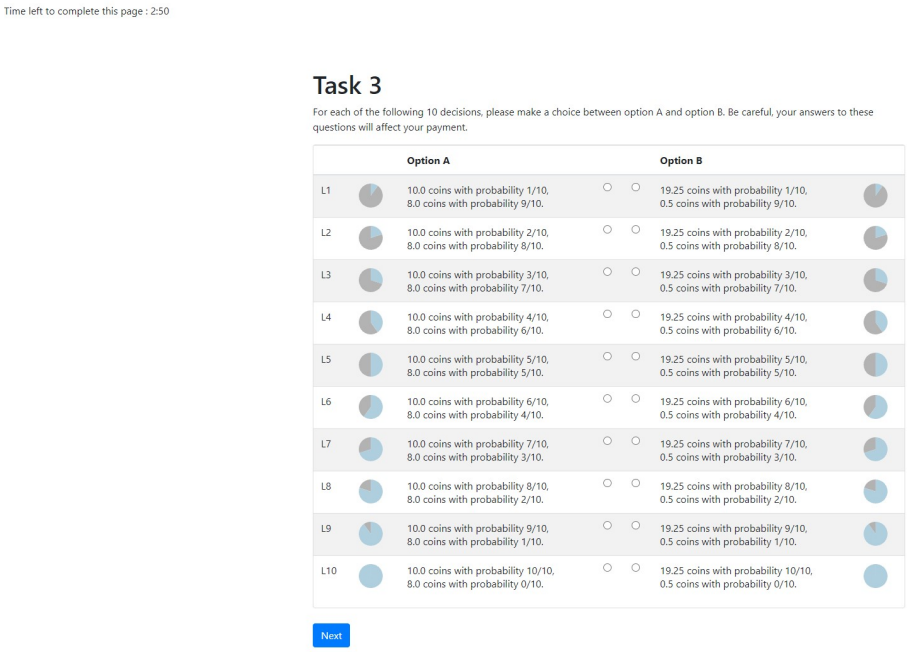


Figure 4.17: Answer to the Risk-Elicitation Task

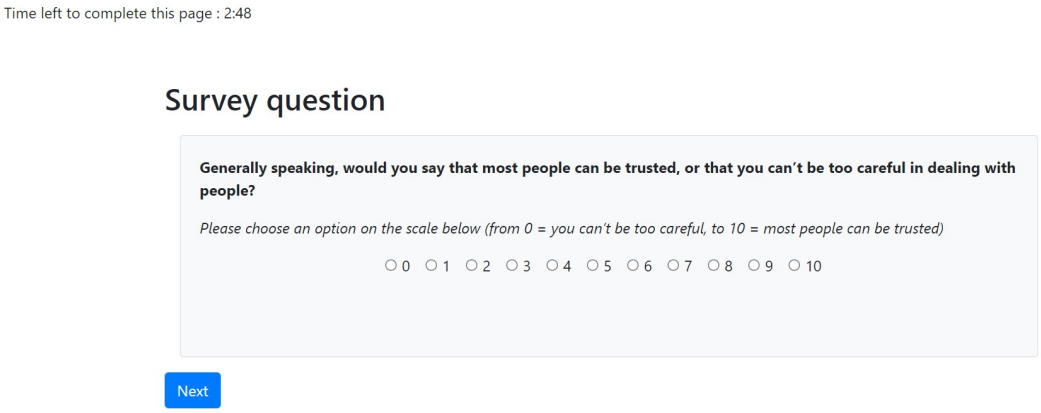


Figure 4.18: Stated Trust Question

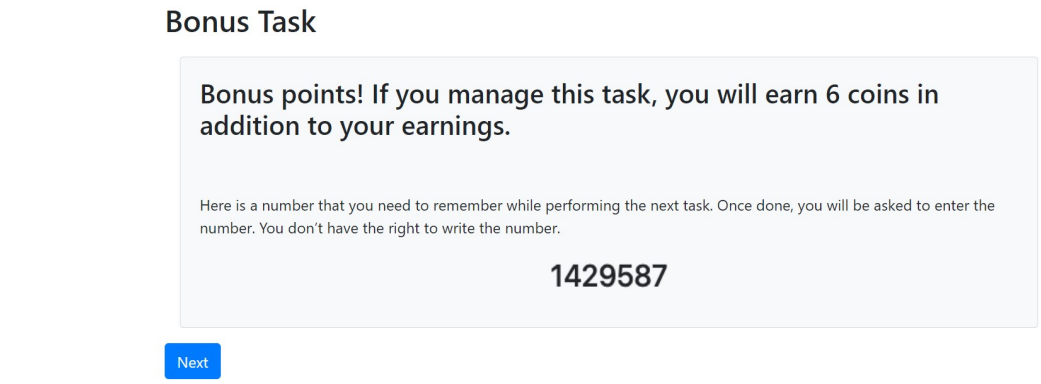


Figure 4.19: Distraction Task, Number Task

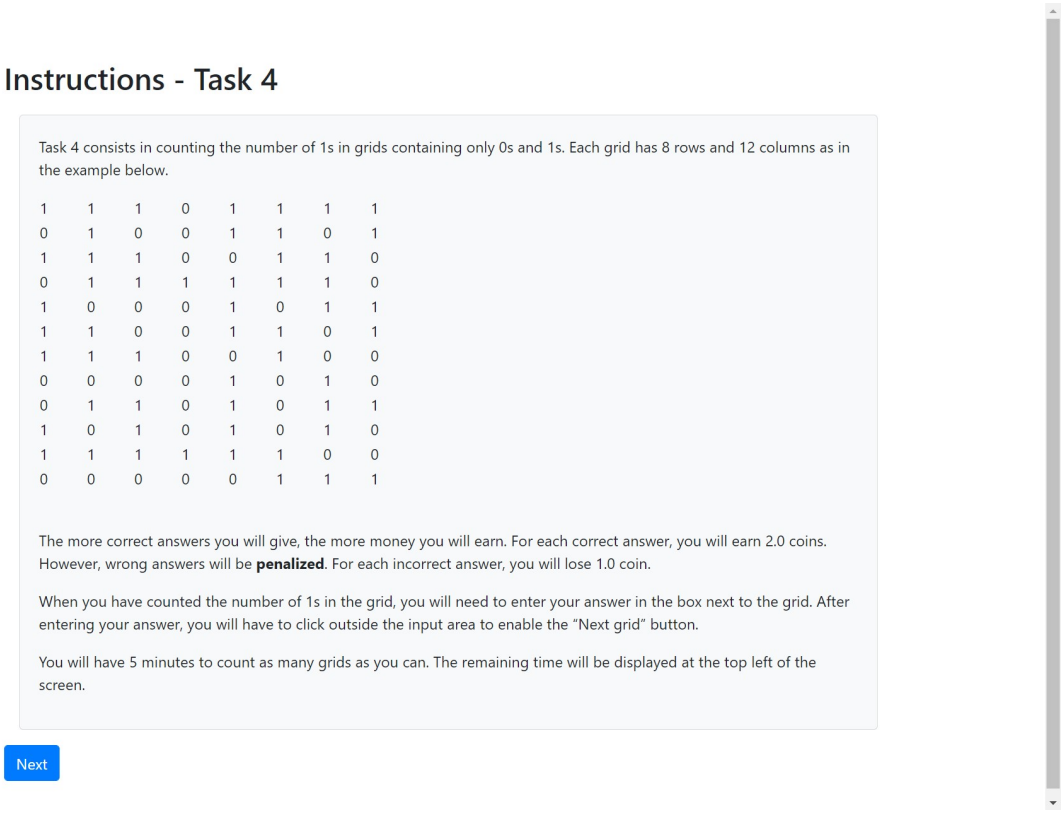


Figure 4.20: Distraction Task, Matrix Instructions

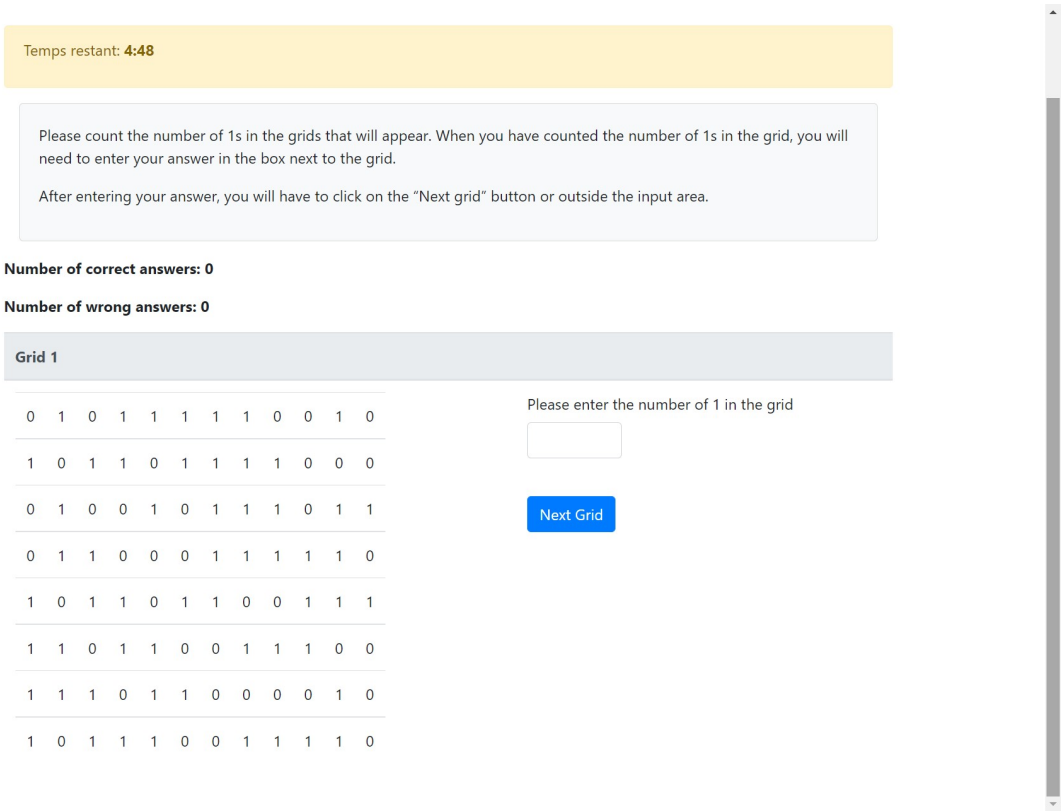


Figure 4.21: Distraction Task, Matrix

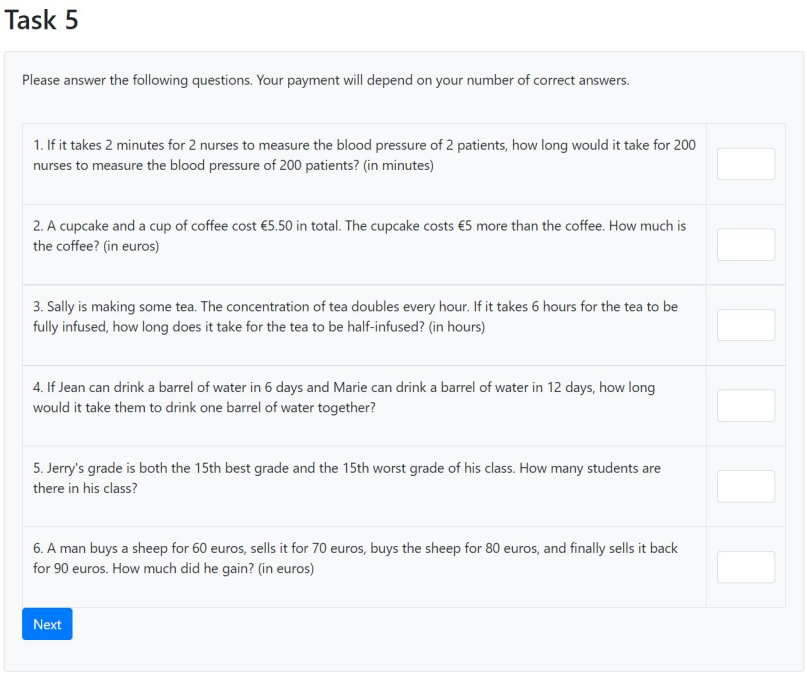


Figure 4.22: CRT Questions

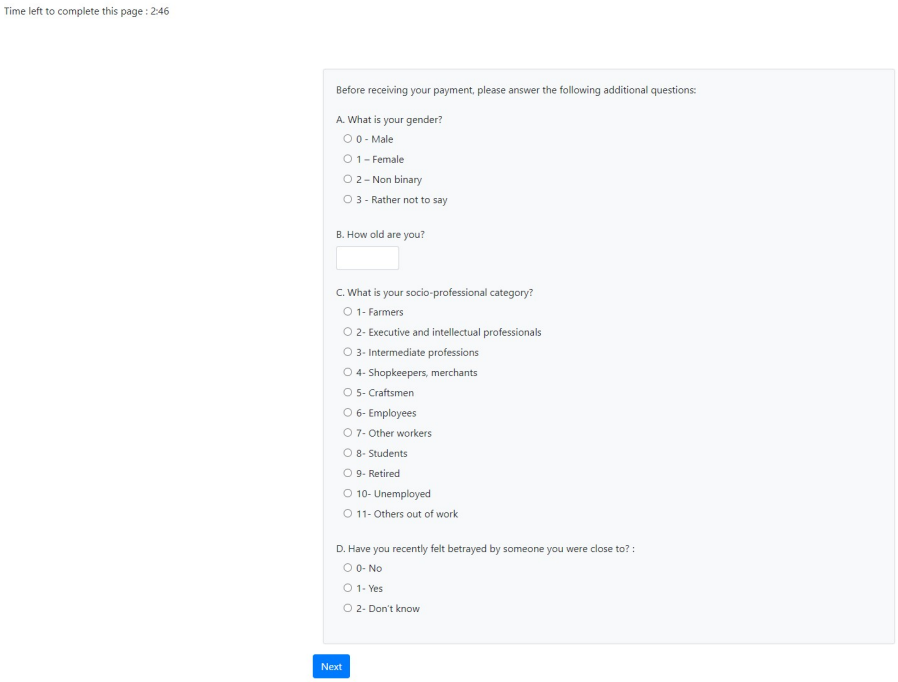


Figure 4.23: Socio-Economic Questions

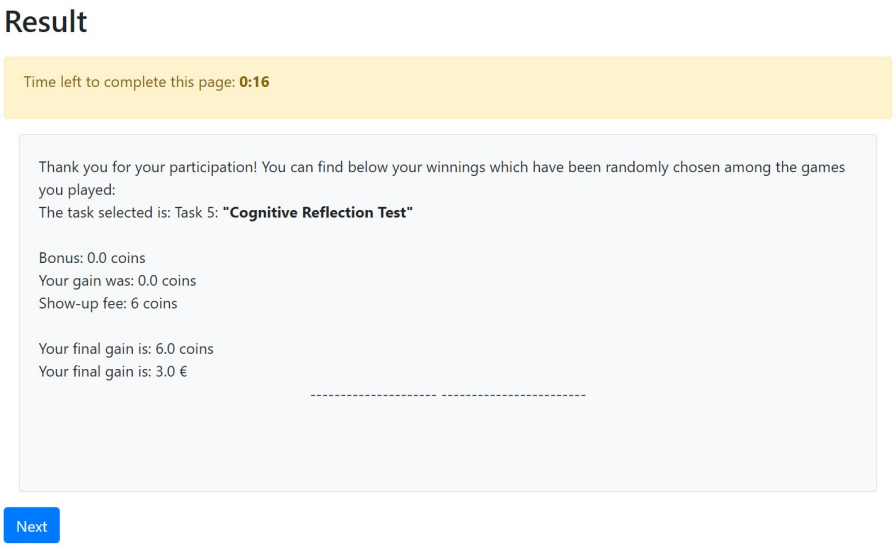


Figure 4.24: Individual Result Presentation

Conclusion

The dissertation contributes to the field of behavioral and experimental economics. Regarding behavioral game theory, we offer insight into the mechanism of self-selection into strategic environments and introduce a new theoretical model. Regarding noise in experimental tasks, we build on the influential work of Gillen, Snowberg, and Yariv and keep on raising awareness on the issue of measurement error in experiments. I would like to end this manuscript by stating three open questions that I consider of prime interest.

A first open question is about the structure of the bounded rationality of economic agents. A pattern seems to emerge from all the studies of the dissertation. When subjects face a task, some of them seem to understand well what they have to do and have a clear idea of how to make a choice at the task at stake (e.g., games, choices between lotteries). Those people may be less noisy in the tasks they face, may want to participate in competitive environments rather than getting a sure payoff, may earn more money than others in experiments. Another fraction of subjects appears more confused and also noisier[1]. Understanding and identifying the link between behavioral biases and noisy behaviors may thus be an exciting question of research.

A second open question concerns the link between bounded rationality and *self-selection* or, more generally, *selection*. An expert in auctions does not have the same characteristics as a real estate agent or a financial trader. They have been selected, by themselves, by the market, or by imitating one of their relatives, and this selection makes them different. This heterogeneity needs to be better understood to model economic decisions. Furthermore, this question is also of particular importance since it is directly related to the external validity of lab experiments.

A last open question is about the complexity of tasks and decisions. Economic agents find some tasks more complex than others. The complexity may stem from the way problems are presented, from the implicated calculations, or the multitude of pieces of information to process. Defining the complexity of economic decision problems as *perceived* by agents may be as difficult as essential in the behavioral economics research agenda.

[1]Similar results are found by Enke and Graeber in their studies on cognitive uncertainty

ECOLE DOCTORALE

Titre : Rationalité Limitée et Bruit : Théorie, Méthodes et Expériences

Mots clés : Rationalité Limitée ; Bruit ; Economie Comportementale ; Economie Expériementale ; Théorie des jeux ; Erreur de mesure

Résumé : La science économique s'intéresse à la façon dont les agents prennent des décisions dans des situations économiques. L'approche standard dans la littérature économique est de supposer que les agents sont parfaitement rationnels. Cependant, l'hypothèse de rationalité parfaite est très restrictive et peine parfois à expliquer le comportement empirique des agents. L'étude de la rationalité limitée des agents est donc une question centrale en économie. Cette thèse contribue à cette question de recherche en économie comportementale et expérimentale, en s'intéressant particulièrement aux comportements dans les environnements stratégiques ainsi qu'aux décisions bruitées (plutôt que purement déterministes). Ce manuscrit est composé de quatre chapitres. Les chapitres 1 et 2 abordent les comportements dans les interactions stratégiques sous différents angles. Les chapitres 3 et 4 s'intéressent à la question du bruit dans les tâches expérimentales. Le chapitre 1 offre un aperçu de la façon dont l'auto-sélection (le fait que les agents décident eux-mêmes s'ils participent à une situation économique) pourrait promouvoir la rationalité dans les environnements stratégiques. L'auto-sélection a rarement été étudiée dans le cadre de la rationalité limitée. Dans la plupart des expériences, une fois assis dans le laboratoire, les sujets sont obligés de participer à toutes les tâches auxquelles ils sont confrontés. Au contraire, dans de nombreuses situations économiques (par exemple les enchères et les marchés financiers), les agents peuvent choisir s'ils veulent participer. À cet égard, il est utile de comprendre les conséquences de l'auto-sélection. Nous évaluons l'évolution des stratégies lorsqu'une étape d'auto-sélection est ajoutée dans des jeux expérimentaux et nous étudions les moteurs potentiels de l'auto-sélection dans les jeux. Le chapitre 2 propose et teste un nouveau modèle de rationalité limitée dans les environnements stratégiques. Le modèle suppose l'existence de deux types de joueurs : les joueurs confus qui ne se forgent aucune croyance et qui jouent donc de manière aléatoire ou naïve, et les joueurs stratégiques qui se forgent des croyances (bruitées) sur la probabilité d'affronter un joueur confus ou stratégique. La stratégie de ces joueurs stratégiques est alors une meilleure réponse à leurs croyances bruitées. Le modèle présente des propriétés théoriques intéressantes. Il est de plus capable de prédire des données expérimentales avec un seul paramètre. Le chapitre 3 calibre l'erreur de mesure dans quatre tâches d'aversion au risque en exploitant un large ensemble d'expériences utilisant un protocole « test-retest » (une même tâche est effectuée plusieurs fois au sein d'une même expérience). Étant donné que l'erreur de mesure peut avoir un impact considérable sur l'analyse statistique, la quantification du bruit dans les mesures expérimentales revêt une importance particulière. Nous examinons également les conséquences de l'erreur de mesure associée à la discrétisation des mesures et aux échantillons de petite taille lors de l'estimation de régressions linéaires. Le chapitre 4 aborde les tâches visant à mesurer la confiance. En utilisant un protocole « test-retest », nous calibrons l'erreur de mesure dans les mesures de confiance standard : le comportement dans un jeu de confiance et les indices de confiance déclaratifs. Nous examinons enfin la façon dont les capacités cognitives et la concentration peuvent avoir un impact sur la quantité de bruit dans ces tâches expérimentales.

Title : Bounded Rationality and Noise : Theory, Methods, and Experiments

Keywords : Bounded Rationality ; Noise ; Behavioral Economics ; Experimental Economics ; Game Theory ; Measurement Error

Abstract : Economists aim at understanding how agents make decisions in economic situations. The standard approach in the economic literature is to assume that agents are perfectly rational. However, the perfect rationality assumption is highly restrictive and does not perfectly capture how agents behave. Studying how bounded is the rationality of economic agents is thus a matter of prime importance. The present dissertation contributes to this research question in behavioral and experimental economics with a particular focus on behaviors in strategic environments and on noisy (rather than purely deterministic) decisions. The dissertation is composed of four chapters. Chapters 1 and 2 address behaviors in strategic interactions with different angles. Chapters 3 and 4 focus on the issue of noise in experimental tasks. Chapter 1 offers insight into how self-selection could promote rationality in strategic environments. Self-selection has seldom been studied under the scope of bounded rationality. In most experiments, once seated in the lab, subjects are forced to participate in all the tasks they faced. On the contrary, in many economic situations in the field (e.g. auctions, financial markets) agents can choose whether they want to participate. In this regard, understanding the consequences of self-selection is valuable. We assess the evolution of strategies when a self-selection stage is added in experimental games and study the potential drivers of self-selection into games. Chapter 2 proposes and tests a new model in behavioral game theory. The model assumes the existence of two types of players : confused players who do not form any beliefs and thus play randomly or naively, and strategic players who best respond to noisy beliefs regarding the probability of facing a confused or a strategic player. The model has interesting theoretical properties and fits experimental data with a single parameter. Chapter 3 calibrates measurement error in four risk-aversion tasks using a large set of existing test-retest experiments. Since measurement error can have a dramatic impact on statistical analysis, quantifying the noise in experimental measures is of particular importance. We also discuss the consequences of measurement error coupled with discrete approximations and small samples when performing linear regressions. Chapter 4 focuses on tasks aiming at eliciting trust. Using a test-retest experiment, we calibrate measurement error in standard trust measures, namely the behavior in a trust game and survey questions. We then discuss how cognitive skills and attention can drive noise in experimental tasks.

Institut Polytechnique de Paris
91120 Palaiseau, France