DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Essays in Experimental Economics“

verfasst von / submitted by

Mag. Alexander Rabas

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 094 140
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Volkswirtschaftslehre
Betreut von / Supervisor:	Univ.-Prof. Dr. Jean-Robert Tyran
Betreut von / Supervisor:	Univ.-Prof. Dr. Wieland Müller

Table of Contents

Introduction

At its core, Economics is about decisions. Whether they are the decisions of individuals, households, firms or countries, economists try to model and understand the behavior of these entities. For a long time, neoclassical Economics applied the idea of a homo economicus, a perfectly rational and narrowly self-interested agent that makes decisions by maximizing its utility. But if all of us acted that way, why do we make choices that we regret later? Why do we also care about others and not only about ourselves? Why do we over- or underestimate our abilities in different contexts when a perfectly rational individual should not have this problem?

To incorporate these ideas into Economics, behavioral economists do not assume that all people behave rationally and only in self-interest. As decisions by economic entities can be seemingly irrational, but need to be explained and understood nonetheless, we incorporate personality, limited rationality and biases to better understand human behavior in markets and games. People are extremely complex in and of themselves, so we consider situations, ideas or potential irrationalities one at a time to get a better understanding of the mystery that is human behavior. Can people overcome their biases by learning? Do men and women have different biases? Do people have consistent preferences or do they change over time? These and more questions are investigated in this dissertation.

In the first part of my thesis, "How to reveal people's preferences: Comparing time consistency and predictive power of multiple price list risk elicitation methods" (with Tamás Csermely, published in the Journal of Risk and Uncertainty, see references), my coauthor and I explore how to measure and elicit risk attitudes. We transform all risk elicitation methods used today into the framework of the most popular risk elicitation method, the multiple price list format. The elicitation methods then differ mostly in two dimensions: forecast accuracy and consistency. Unlike the idea of a perfectly rational individual, people are often inconsistent in their choices over a short time frame, so some methods produce more stable results than others. As accuracy and consistency are necessary aspects of all methods of elicitation, finding out which methods yield more stable and more accurate results to mitigate bounded rationality is of the utmost importance to quantify human behavior. A derivation of the well-known risk elicitation method by Holt and Laury (2002), devised by Drichoutis and Lusk (2012), emerges as the superior method in both dimensions, and we recommend to use this method in future research.

The second part of my thesis, "The Sequential Hotelling Game - Slowly learning the Equilibrium Outcome" (single-authored), addresses decisions in a gametheoretical and experimental context that can be used to model political elections. The game in question is a three-player variant of Hotelling's (1929) influential locational choice model, where entry is sequential instead of simultaneous. Martin Osborne and Amoz Kats offer a conjecture for this game, namely that the first and the last player choose as their location the median, and all other players do nothing; I prove this conjecture for the special case of three players. Because of the nature of the equlibrium, this game is then related to Duverger's Law (1959), as only two players enter the game. Although a clear prediction exists how people should behave in this context, initial decisions by subjects are completely different from theory. In particular the realization that sometimes the best course of action is to do nothing seems hard to grasp. However, it is interesting to see that even when gametheoretical predictions do not hold initially, Duverger's Law is robust to deviations from theory also in this context, as initial play consists of the first two players entering the game and the third doing nothing. With enough time people are able to learn the optimal way to make decisions, so while results change massively depending on the timeframe, behaving rationally and overcoming biases can be learned by repetition.

The third part of my thesis, "Underconfident Women Earn Less - A Virtual Lab Approach" (with Jean-Robert Tyran and Rupert Sausgruber), analyzes the impact of beliefs on occupational choices in the labor market for men and women in a controlled, experimental setting. There are many reasons why men and women might choose different jobs, and depending on the particular reason those choices might be efficient or not. If, for instance, men and women have different preferences over job characteristics or social preferences, the fact that they sort into different occupations can be seen as efficient, as individuals act as utility maximizers. However, we show that confidence into one's own skills is an important determinant of sorting into jobs, and that men and women make systematically different choices: Highly productive women who are underconfident in their abilities tend to sort into jobs that are unprofitable for them, which is inefficient and some of women's talents are lost. This result then also carries over to incomes earned in the Danish labor market, as these productive women who we find to be underconfident earn less than otherwise similar women who are not underconfident. This indicates that to understand human behavior, a multitude of behavioral dimensions, including differences in gender, must be analyzed and understood to predict and model human decision making.

I hope that through the insights my coauthors and I added to the literature, we were able to, at least in part, answer some of the fundamental questions above. Or, as the Nobel laureate Herbert A. Simon says: "Anything that gives us new knowledge gives us an opportunity to be more rational."

References

Csermely, T. & Rabas, A. (2016). How to reveal peoples preferences: Comparing time consistency and predictive power of multiple price list risk elicitation methods. *Journal of Risk and Uncertainty*, 53(2-3), 107-136.

Drichoutis, A. & Lusk, J. (2012). Risk preference elicitation without the confounding effect of probability weighting. Working paper. Munich Personal RePEc Archive, No. 37776.

Duverger, Maurice (1954). *Political Parties*. London: Methuen.

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

Hotelling, H. (1929). Stability in Competition. *Economic Journal* 39, 41-57.

How to reveal people's preferences: Comparing time consistency and predictive power of multiple price list risk elicitation methods

Tamás Csermely[*] Alexander Rabas[†]

June 23, 2017[‡]

Abstract

The question of how to measure and classify people's risk preferences is of substantial importance in the field of economics. Inspired by the multitude of ways used to elicit risk preferences, we conduct a holistic investigation of the most prevalent method, the multiple price list (MPL) and its derivations. In our experiment, we find that revealed preferences differ under various versions of MPLs as well as yield unstable results within a 30-minute time frame. We determine the most stable elicitation method with the highest forecast accuracy by using multiple measures of within-method consistency and by using behavior in two economically relevant games as benchmarks. A derivation of the well-known method by Holt and Laury (2002), where the highest payoff is varied instead of probabilities, emerges as the best MPL method in both dimensions. As we pinpoint each MPL characteristic's effect on the revealed preference and its consistency, our results have implications for preference elicitation procedures in general.

Keywords: Risk, Multiple Price List, MPL, Revealed Preferences

JEL Classification: C91 · D81

[*] Vienna University of Economics and Business, Institute for Public Sector Economics

[†] University of Vienna, Doctoral School of Economics Vienna, Department of Economics
alexander.rabas@gmail.com

[‡] Published in the Journal of Risk and Uncertainty (2016), 53(2-3), 107-136.

1 Introduction

Risk is a fundamental concept that affects human behavior and decisions in many real-life situations. Whether a person wants to invest in the stock market, tries to select the best health insurance or just wants to cross the street, he/she will face risky decisions every day. Therefore, risk attitudes are of high importance for decisions in many economics-related contexts. A multitude of studies elicit risk preferences in order to control for risk attitudes, as it is clear that they might play a relevant role in explaining results – e.g. de Véricourt et al. (2013) in the newsvendor setting, Murnighan et al. (1988) in bargaining, Beck (1994) in redistribution or Tanaka et al. (2010) in linking experimental data to household income, to name just a few. Moreover, several papers try to shed light on the causes of risk-seeking and risk-averse behavior in the general population with laboratory (Harrison and Rutström, 2008), internet (von Gaudecker et al., 2011) and field experiments (Andersson et al., 2016; Harrison et al., 2007). Since the seminal papers by Holt and Laury (2002, 2005), approximately 20 methods have been published which provide alternatives to elicit risk preferences. They differ from each other in terms of the varied parameters, representation and framing. Many of these risk elicitation methods have the same theoretical foundation and therefore claim to measure the same parameter – a subject's "true" risk preference. However, there are significant differences in results depending on the method used, as an increasing amount of evidence suggests. It follows that if someone's revealed preference is dependent on the measurement method used, scientific results and real-world conclusions might be biased and misleading.

As far as existing comparison studies are concerned, they usually compare two methods with each other and often use different stakes, parameters, framing, representation, etc., which makes their results hardly comparable. Our paper complements existing experimental literature by making the following contribution: Taking the method by Holt and Laury (2002) as a basis, we conduct a comprehensive comparison of the multiple price list (MPL) versions of risk elicitation methods by classifying all methods into nine categories. To the best of our knowledge, no investigation – including various measures of between- and within-method consistency – has ever been conducted in the literature that incorporates such a high number of methods. To isolate the effect of different methods, we consistently use the MPL representation and calibrate the risk intervals to be the same for each method assuming expected utility theory (EUT) and constant relative risk aversion (CRRA), while also keeping the risk-neutral expected payoff of each method constant and employing a within-subject design. Moreover, our design allows us to investigate whether differences across methods can be reconciled by assuming different functional forms documented in the literature such as constant absolute risk aversion (CARA), decreasing relative risk aversion (DRRA), decreasing absolute risk aversion (DARA), increasing relative risk aversion (IRRA) and increasing absolute risk aversion (IARA). Additionally, we extend our analysis to incorporate EUT with probability weighting and also to incorporate prospect theory (PT) and cumulative prospect theory (CPT).

We investigate the within-method consistency of each method by comparing the differences in subjects' initial and repeated decisions within the same MPL method. Moreover, we assess methods' self-perceived complexity and shed light on differences and similarities between methods. In the end, we provide suggestions for which specific MPL representation to use by comparing our results to decisions in two benchmark games that resemble real-life settings: investments in capital markets and auctions. Therefore, we analyze the methods along two dimensions, robustness and predictive power, and determine which properties of particular methods drive risk attitude and its consistency.

We find that a particular modification of the method by Holt and Laury (2002) derived by Drichoutis and Lusk (2012, 2016) has the highest predictive power in investment settings both according to the OLS regression and Spearman rank correlation. In addition, specific methods devised by Bruner (2009) also perform relatively well in these analyses. However, the method by Drichoutis and Lusk (2012, 2016) clearly outperforms the other methods in terms of within-method consistency and is perceived as relatively simple – in the end, our study provides the recommendation for researchers to implement this method when measuring risk attitudes using an MPL framework. Moreover, our results remain qualitatively the same if we relax our assumption on the risk aversion function, or if we take probability weighting or alternative theories such as prospect theory or cumulative prospect theory into account.

1.1 Multiple Price Lists Explained

Incentivized risk preference elicitation methods aim to quantify subjects' risk perceptions based on their revealed preferences. We present nine methods in a unified structure – the commonly used MPL format – to our subjects, taking one of the most cited methods as a basis: Holt

Table 1: *Risk parameter intervals (Holt/Laury)*

Interpretation by Holt/Laury (2002)	Switching Point	Risk Parameter Interval
highly risk loving	1	$\rho \leq -0.95$
very risk loving	2	$-0.95 < \rho \leq -0.49$
risk loving	3	$-0.49 < \rho \leq -0.15$
risk neutral	4	$-0.15 < \rho \leq 0.15$
slightly risk averse	5	$0.15 < \rho \leq 0.41$
risk averse	6	$0.41 < \rho \leq 0.68$
very risk averse	7	$0.68 < \rho \leq 0.97$
highly risk averse	8	$0.97 < \rho \leq 1.37$
stay in bed	Never	$\rho > 1.37$

Notes: This table indicates the mapping from a subject's chosen switching point into the resulting risk parameter intervals in each method; the leftmost column contains the interpretation of the risk intervals; "Never" means a subject prefers the option "Left" in each row

and Laury (2002). The MPL table structure is as follows: Each table has multiple rows, and in each row all subjects face a lottery (two columns) on one side of the table, and a lottery or a certain payoff (one or two columns) on the other side, depending on the particular method. Then, from row to row, one or more of the parameters change. The methods differ from each other by the parameter which is changing. As the options on the right side become strictly more attractive from row to row, a subject indicates the row where he/she wants to switch from the left option to the right option. This switching point then gives us an interval for a subject's risk preference parameter according to Table 1[1], assuming EUT and CRRA[2].

[1] To ease comparison to existing studies, we used exactly the same coefficient intervals as Holt and Laury (2002).

[2] $u(c) = \frac{c^{1-\rho}}{1-\rho}$

Note that several other representations of risk elicitation methods exist besides the MPL such as the bisection method (Andersen et al., 2006), the trade-off method (Wakker and Deneffe, 1996), questionnaire-based methods (Weber et al., 2002), willingness-to-pay (Hey et al., 2009), etc., but the MPL is favored because of its common usage. Andersen et al. (2006) consider that the main advantage of the MPL format is that it is transparent to subjects and it provides simple incentives for truthful preference revelation. They additionally list its simplicity and the little time it takes as further benefits. As far as the specific risk elicitation method in the MPL framework designed by Holt and Laury (2002) is concerned, it has proven itself numerous times in providing explanations for several phenomena such as behavior in 2x2 games (Goeree et al., 2003), market settings (Fellner and Maciejovsky, 2007), smoking, heavy drinking, being overweight or obese (Anderson and Mellor, 2008), consumption practices (Lusk and Coble, 2005) and many others.

Early studies document violations of EUT under risky decision making and provide suggestions how these differences can be reconciled (Bleichrodt et al., 2001). In addition, recent studies (Tanaka et al., 2010; Bocqueho et al., 2014) document potential empirical support for prospect theory (PT, Kahneman and Tversky, 1979)[3] when it comes to risk attitudes: Harrison et al. (2010) found that PT describes behavior of half of their sample best. There is also evidence that subjective probability weighting (PW) (Quiggin, 1982) should be taken into account. In addition, cumulative prospect theory (CPT) – PT combined with PW (Tversky and Kahneman, 1992) – might also be a candidate that can explain the documented anomalies under EUT. Wakker (2010) provides an extensive review on risk under PT.

We justify using CRRA as Wakker (2008) claims that it is the most commonly postulated assumption among economists. Most recently, Chiappori and Paiella (2011) provide evidence on the validity of this assumption in economic-financial decisions.[4] Nevertheless, alternative functional forms have been introduced, e.g. CARA[5] (Pratt, 1964). It was also questioned whether social status – and mostly the role of wealth or income – might shape risk attitude, which would lead to functions which are increasing or decreasing in these factors such as

[3] $u(c) = \begin{cases} c^{\alpha} & \text{if } c \geq 0 \\ -\lambda(-c)^{\beta} & \text{if } c < 0 \end{cases}$

[4] Note that this approach is also popular among economists due to its computational ease: The vast majority of economic experiments assumes CRRA as well, which makes our results comparable to theirs.

[5] $u(c) = \frac{-e^{\rho c}}{\rho}$

IRRA and DRRA (Andersen et al., 2012)[6] or IARA and DARA (Saha, 1993).[7] A review of these functions is provided by Levy (1994). In our robustness analysis, we relax our original assumptions on EUT and CRRA and incorporate all of the above mentioned alternative theories and functional forms. Note that even though we calibrated our parameters to accomodate EUT and CRRA, one is still able to calculate the risk parameter ρ using the aforementioned alternative specifications.[8]

We group our aforementioned nine risk elicitation methods into two categories:

1. The standard gamble methods (SG methods), where on one side of the table there is always a 100% chance of getting a particular certain payoff and on the other side there is a lottery.

2. The paired gamble methods (PG methods), with lotteries on both sides.

We therefore primarily conduct a comparison of different MPL risk elicitation methods. What we do not claim, however, is that the method devised by Holt/Laury (2002) (or MPL in general) is the most fitting to measure people's risk preferences – we strive to give a recommendation to researchers who already intend to use Holt/Laury (2002) in their studies, and provide a better alternative that shares its attributes with the original MPL design.

It should be mentioned that there is an alternative interpretation of our study: The different MPL methods can also be conceived as a mapping of existing risk elicitation methods (from other frameworks) to the MPL space. Several methods exist where the risk elicitation task is provided in a framed context – such as pumping a balloon until it blows (Lejuez et al., 2002) or avoiding a bomb explosion (Crosetto and Filippin, 2013). Similarly, some methods differ due to the representation of probabilities with percentages (Holt and Laury, 2002) or charts (Andreoni and Harbaugh, 2010). All these methods can be displayed with different MPLs by showing the probabilities and the corresponding payoffs in a table format. We provide a complete classification of these methods in the Literature Review section.

[6] $u(c, W) = [(\omega W^r + c^r)^{(1-\rho)/r}]/(1-\rho)$

[7] $u(c, W) = \frac{-e^{-\rho r(c+W\omega)}}{\rho}$

[8] This implies that the same switching point in two methods does not yield the same risk parameter estimate under different specifications, but these estimates are still directly comparable according to theory, as they claim to measure a subject's underlying risk attitudes ceteris paribus.

Up to now, different risk elicitation methods were compared by keeping the original designs, but this approach comes at a price: As the methods differ in many dimensions, any differences found can be attributed to any of those particular characteristics. Our approach can be understood as a way to make all risk elicitation methods as similar as possible, with the drawback of losing the direct connection to the original representation. This paper should therefore primarily be seen as a comparison of different MPL risk elicitation methods, and the resulting comparison of existing risk elicitation methods by mapping them into the same space is only reported for the sake of completeness.

1.2 Literature Review

We will now discuss the different methods in greater detail and how they are embedded in the literature, if at all. Table 2 provides a summary of the exact parameter that is changing across methods.[9]

A Standard Gamble Methods

Among the SG methods, there are four parameters that can be changed: The sure payoff ($sure$), the high payoff of the lottery ($high$), the low payoff of the lottery (low) or the probability of getting the high payoff (p) (or the probability of getting the low payoff $(1-p)$, respectively). The parameters must of course be chosen in such a way that $high > sure > low$ always holds. For instance, we denote the SG method where the low payoff is changing by "SGlow", the SG method with the varying certainty equivalent by "SGsure" or the standard gamble method where the probabilities are changing as "SGp".

Binswanger (1980) introduced a method (SGall) where only one of the options has a certainty equivalent. The other options consist of lotteries where the probabilities are fixed at 50-50, but both the high and the low payoff are changing. Cohen et al. (1987) used risk elicitation tasks in which probabilities and lottery outcomes were held constant and only the certainty equivalent was varied. These methods have later been redesigned and presented in a more sophisticated format as a single choice task by Eckel and Grossman (2002, 2008).

[9] For a complete list of all methods with the corresponding parameter values (as presented to subjects), refer to the Online Resource.

Table 2: *Method overview*

	What is changing?			
Method	Probability	Highest Payoff	Lowest Payoff	Sure Payoff
SGp	**yes**	no	no	no
SGhigh	no	**yes**	no	no
SGlow	no	no	**yes**	no
SGsure	no	no	no	**yes**
SGall	no	**yes**	**yes**	**yes**
PGp	**yes**	no	no	NA
PGhigh	no	**yes**	no	NA
PGlow	no	no	**yes**	NA
PGall	**yes**	**yes**	**yes**	NA

Notes: This table indicates which parameters change from row to row in each method, where SG stands for "standard gamble" and PG stands for "paired gamble".

A recent investigation by Abdellaoui et al. (2011) presents a similar method (SGsure method) in an MPL format with 50-50 probabilities. Bruner (2009) presents a particular certainty equivalent method, where the certainty equivalent and the lottery outcomes are held constant, but the corresponding probabilities of the lotteries are changing (SGp method). Additionally, Bruner (2009) introduces another method where only the potential high outcomes of lotteries vary (SGhigh method). Although not present in the literature, we chose to include a method where the potential low outcome varies for reasons of completeness (SGlow method).[10]

B Paired-Gamble Methods

Holt and Laury (2002, 2005) introduced the most-cited elicitation method under EUT up to now, where potential outcomes are held constant and the respective probabilities change (PGp). Drichoutis and Lusk (2012, 2016) suggest a similar framework where the outcomes of different lotteries change while the probabilities are held constant. We differentiate these methods further into PGhigh and PGlow depending on whether the high or the low outcome is varied in the

[10] For examples, see Tables $13-17$ in the Online Resource, which correspond to the SG methods.

MPL. Additionally, the PGall method varies both the probabilities and the potential earnings at the same time.

Many risk elicitation tasks used in the literature fit into the framework of choosing between different lotteries. Sabater-Grande and Georgantzis (2002) provide ten discrete options with different probabilities and outcomes to choose from. Lejuez et al. (2002) introduce the Balloon Analogue Risk Task where subjects could pump up a balloon, and their earnings depend on the final size of the balloon. The larger the balloon gets, the more likely it will explode, in which case the subject earns nothing. Visschers et al. (2009) and Andreoni and Harbaugh (2010) use a pie chart for probabilities and a slider for outcomes to visualize a similar trade-off effect in their experiment. Crosetto and Filippin (2013) present their Bomb Risk Elicitation Task with an interesting framing which quantifies the aforementioned trade-off between probability and potential earnings with the help of a bomb explosion.[11]

C Questionnaire Methods

In addition to the MPL methods, we chose to also incorporate questionnaire risk elicitation methods into our study. Several methods have been introduced that evaluate risk preferences with non-incentivized survey-based methods, and the questions and the methodology they use are very similar. The most recently published ones include the question from the German Socio-Economic Panel Study (Dohmen et al., 2011) or the Domain-Specific Risk-Taking Scale (DOSPERT) by Blais and Weber (2006). For a more detailed description, see the last paragraph of Section 2.

D Comparison Studies

The question arises of which method to use if there is such a large variety of risk elicitation methods and whether they lead to the same results. Comparison studies exist, but the majority compare two methods with each other, and thus their scope is limited. The question of within-method consistency has been addressed by some papers: Harrison et al. (2005) document high re-test stability of the method introduced by Holt and Laury (2002, PGp). Andersen et al. (2008b) test consistency of the PGp (2002) method within a 17-month time frame. They

[11] For examples, see Tables $18 - 21$ in the Online Resource, which correspond to the PG methods.

find some variation in risk attitudes over time, but do not detect a general tendency for risk attitudes to increase or decrease. This result was confirmed in Andersen et al. (2008a). Yet there is a gap in the academic literature on the time stability of different methods and their representation that we are eager to fill.

Interestingly, more work has been done on the field of between-method consistency. Fausti and Gillespie (2000) compare risk preference elicitation methods with hypothetical questions using results from a mail survey. Isaac and James (2000) conclude that risk attitudes and relative ranking of subjects is different in the Becker-DeGroot-Marschak procedure and in the first-price sealed-bid auction setting. Berg et al. (2005) confirm that assessment of risk preferences varies generally across institutions in auction settings. In another comparison study, Bruner (2009) shows that changing the probabilities versus varying the payoffs leads to different levels of risk aversion in the PG tasks. Moreover, Dave et al. (2010) conclude that subjects show different degrees of risk aversion in the Holt and Laury (2002, PGp) and in the Eckel and Grossman (2008, SGall) task. Their results were confirmed by Reynaud and Couture (2012) who used farmers as the subject pool in a field experiment. Bleichrodt (2002) argues that a potential reason for these differences might be attributed to the fact that the original method by Eckel and Grossman (2008) does not cover the risk seeking domain, which can be included with the slight modification we made when incorporating this method. Dulleck et al. (2015) test the method devised by Andreoni and Harbaugh (2010) using a graphical representation against the PGp and describe both a surprisingly high level of within- and inter-method inconsistency. Drichoutis and Lusk (2012, 2016) compare the PGp method to a modified version of it where probabilities are held constant. Their analysis reveals that the elicited risk preferences differ from each other both at the individual and at the aggregate level. Most recently, Crosetto and Filippin (2016) compare four risk preference elicitation methods with their original representation and framing and confirm the relatively high instability across methods.

In parallel, a debate among survey-based and incentivized preference elicitation methods emerged which were present since the survey on questionnaire-based risk elicitation methods by Farquhar (1984). Eckel and Grossman (2002) conclude that non-incentivized survey-based methods provide misleading conclusions for incentivized real-world settings. In line with this finding, Anderson and Mellor (2009) claim that non-salient survey-based elicitation methods and the PGp method yield different results. On the contrary, Lönnqvist et al. (2015) provide evidence that the survey-based measure, which Dohmen et al. (2011) had implemented, explains decisions in the trust game better than the SGsure task. Charness and Viceisza (2016) provide evidence from developing countries that hypothetical willingness-to-risk questions and the PGp task deliver deviating results.

E Further Considerations

A recent stream of literature broadens the horizon of investigation to theoretical aspects of elicitation methods: Weber et al. (2002) show that people have different risk attitudes in various fields of life, thus risk preferences seem to be domain-specific. Lönnqvist et al. (2015) document no significant connection between the HLp task and personality traits. Dohmen et al. (2010) document a connection between risk preferences and cognitive ability, which was questioned by Andersson et al. (2016). Hey et al. (2009) investigate noise and bias under four different elicitation procedures and emphasize that elicitation methods should be regarded as strongly context specific measures. Harrison and Rutström (2008) provide an overview and a broader summary of elicitation methods under laboratory conditions, whereas Charness et al. (2013) survey several risk preference elicitation methods based on their advantages and disadvantages.

In addition, there is evidence that framing and representation matters. Wilkinson and Wills (2005) advised against using pie charts showing probabilities and payoffs as human beings are not good at estimating angles. Hershey et al. (1982) identify important sources of bias to be taken into account and pitfalls to avoid when designing elicitation tasks. Most importantly, these include task framing, differences between the gain and loss domains and the variation of outcome and probability levels. Von Gaudecker et al. (2008) show that the same risk elicitation methods for the same subjects deliver different results when using different frameworks – e.g. multiple price list, trade-off method, ordered lotteries, graphical chart representation, etc. This procedural indifference was confirmed by Attema and Brouwer (2013) as well, which implies that risk preferences on an individual level are susceptible to the representation and framing used.

The previous paragraphs lead us to the conclusion that methods should be compared to each other by using the same representation and format. This justifies our decision to compare them using the standard MPL framework which guarantees that the differences cannot be attributed to the different framing and representation of elicitation tasks. However, this comes

Table 3: *Link between MPL representation and literature*

Method	Corresponding Literature
SGp	Bruner (2009)
SGhigh	Bruner (2009)
SGlow	
SGsure	Cohen et al. (1987), Abdellaoui et al. (2011)
SGall	Binswanger (1980), Eckel and Grossman (2008)
PGp	Holt and Laury (2002), Holt and Laury (2005)
PGhigh	Drichoutis and Lusk (2012, 2016)
PGlow	Drichoutis and Lusk (2012, 2016)
PGall	Sabater-Grande and Georgantzis (2002), Lejuez et al. (2002), Andreoni and Harbaugh (2010), Crosetto and Filippin (2013)
Questionnaire	Weber et al. (2002), Dohmen et al. (2011)

Notes: On the left, this table lists all MPL and questionnaire methods, and on the right the corresponding literature.

at the price that we had to change some of the methods slightly, which implies that they are not exactly the same as their originally published versions. We certainly do not claim that the MPL is the only valid framework, but our choice for it seems justified by its common usage and relative simplicity. We consider a future investigation using a different representation technique as a potentially interesting addition. Also, we emphasize that the differences in our results exist among the MPL representations of the methods and they can only be generalized to the original methods to a very limited extent. See Table 3 for an overview of the link between the MPL representation and the particular method that was published originally, and Table 12 in Appendix section A.2, where we compared the results from our MPL methods to the results in previous studies – most of the studies deliver significantly different results to the risk parameters measured in our study. This is not surprising given the considerations in Sections 1.2.D and 1.2.E, as we map all methods to the MPL space. Furthermore, risk elicitation methods are very noisy in general. For example the same method with the same representation delivers significantly different results in Crosetto and Filippin (2013) and Crosetto and Filippin (2016).

2 Design

We provide a laboratory experiment to compare different MPL risk elicitation methods. Subjects answered the risk elicitation questions first. Then, benchmark games were presented to them to gauge predictive power, which was followed by a non-incentivized questionnaire. We will provide a detailed description on the exact procedures of each part in the later paragraphs.

We conducted ten sessions at the Vienna Center for Experimental Economics (VCEE) with 96 subjects.[12] Sessions lasted about 2 hours, with a range of earnings between €3 and €50, amounting to an average payment of €20.78 with a standard deviation of €10.1. We calibrated these payments similarly to previous studies (e.g. Bruner, 2009 or Abdellaoui et al., 2011, among others). Average earnings were about €9.5 in the risk task and about €8.3 in the benchmark games plus a €3.00 show-up fee. Harrison et al. (2009) provide evidence that the existence of a show-up fee could lead to an elevated level of risk aversion in the subject pool. In our experiment, this moderate show-up fee was only pointed out to the subjects after making their decisions in the risk elicitation methods and the benchmark games. Thus, it could not have distorted their preferences. The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007), and ORSEE (Greiner, 2015) was used for recruiting subjects.

We employed a within-subject design, meaning that each subject took decisions in each and every task as in other comparison studies (Eckel and Grossman, 2008; Crosetto and Filippin, 2016). This property rules out that the methods differ due to heterogeneity between subjects, but it comes with the drawback that methods which were encountered later might deliver more noisy or different results due to fatigue or other factors, as the answer to a particular method could also be a function of previously seen MPLs. Consequently, we included the order in which a method appeared in all regressions as controls wherever possible, compared the variance in earlier and latter methods and tested for order effects; no significant effects were found.[13] To avoid biases, a random number generator determined the order of methods for each subject

[12] One subject has been excluded from our subject pool after repeatedly being unable to answer the control questions correctly.

[13] See Table 11 in appendix section A.1.

separately in the beginning of each session.[14]

After receiving instructions on screen and in written form, subjects went through the nine incentivized risk elicitation methods. In order to avoid potential incentive effects mentioned by Holt and Laury (2002), the expected earnings for a risk-neutral individual were equal in every method. Furthermore, to avoid potential biases due to the different reactions to gains and losses (Hershey et al., 1982), each of our lotteries is set in the gains domain. Andersen et al. (2006) confirmed previous evidence (Poulton, 1989) that there is a slight tendency of anchoring and choosing a switching point around the middle for risk elicitation tasks. In order to counteract anchoring and one-directional distortion of preferences as a consequence of this unaviodable pull-to-center effect, each risk elicitation task appeared randomly either top-down or bottom-up. Depending on randomization, out of nine potential switching opportunities the fourth or the sixth option were the risk-neutral switching points.[15]

Subjects also had the opportunity to look at their given answer and modify it right after each decision if they wished to do so. After making a decision in each method, we asked subjects the following question: "On a scale from 1 to 10, how difficult was it for you to make a decision in the previous setting?" With this question we assessed self-perceived complexity of the tasks, since there is evidence in the literature (Mador et al., 2000) that subjects make noisier decisions if the complexity of a lottery increases, and therefore a less complex method is preferred. Moreover, Dave et al. (2010) outline the trade-offs between noise, accuracy and subjects' mathematics skills. They suggest that it is a good strategy to make MPL tasks simpler for subjects. In this spirit, we asked our subjects to indicate the row in which they switched from the "LEFT" column to the "RIGHT" column, thereby enforcing a single switching point (SSP). Using this framework, subjects were not required to make a decision for each and every row in every method, which would have meant more than 100 monotonous, repetitive binary choices throughout the experiment. Additionally, this approach ensures that the subjects were guaranteed to give answers without preference reversals. We consider this option more viable than accepting multiple switching points – thus allowing inconsistent choices – and using the total number of "safe" choices to determine a subject's risk coefficient interval. The SSP has

[14] Each subject encountered the methods in a unique random order and each order was used only once in the entire experiment.

[15] An example for the difference between the top-down and bottom-up representation can be found in Table 22 in the Online Resource.

been used several times, e.g. Gonzalez and Wu (1999) or Jacobson and Petrie (2009).

By enforcing a SSP, we faced a trade-off between potential boredom and the non-detection of people with inconsistent preferences. Furthermore, some of the reported within-method instability might stem from "fat preferences" or indifference between two or more options. However, the SSP can further be justified in that only a small proportion of subjects expressed multiple switching points in earlier studies,[16] so this design choice is highly unlikely to drive our results.

In order to test within-method consistency, three of the nine methods were randomly chosen and presented to subjects again, without telling them that they had already encountered that particular method.[17] Repeating all methods was not feasible due to fatigue concerns, as the experiment is already quite long. This approach allows us to test both within-method and inter-method consistency. The modification of subjects' answers was allowed here once as well. The perceived complexity of tasks was also elicited again.

Control questions were used for the preference elicitation methods and for each benchmark game in order to verify that subjects understood the task they were about to perform.[18] Subjects had to answer them correctly in order to participate in the experiment.

We incorporated the random lottery incentive system emphasized by Cubitt et al. (1998). Thus, the computer chose one of the twelve risk preference methods and one of the eight rows within that particular method on a random basis to be payoff-relevant. Additionally, one of the three benchmark games was chosen to be payoff-relevant as well. This random lottery incentive system helps keep the costs at a reasonable level while having similarly sized stakes (than e.g. Bruner, 2009) or even larger stakes (than e.g. Holt and Laury, 2002 or Harrison et al., 2007) for the elicitation tasks compared to previous studies, while mitigating potential income effects. Nevertheless, we note that the random lottery incentive system might be a potential caveat in our study, since Cox et al. (2015) document somewhat different behavior under various payment mechanisms.

[16] E.g. 8.5% in Dave et al. (2010) and 6.6% in Holt and Laury (2002)

[17] The exact number of occasions a particular method was encountered a second time is roughly equal across methods: PGhigh(33), PGlow(25), PGp(26), PGall(39), SGhigh(30), SGsure(33), SGlow(27), SGp(40), SGall(35).

[18] See the Online Resource for the exact text of these questions.

As far as hedging behavior is concerned, Blanco et al. (2010) provide evidence that hedging and the corresponding biased beliefs and actions can only be problematic if the hedging opportunities are highly transparent. Taking this consideration into account, we provided feedback on the outcome of the risk elicitation tasks only at the end of the experiment. Thus, it was not possible for subjects to create a portfolio and use hedging behavior over different parts of the experiment.

On top of the risk elicitation tasks, we used three benchmark games resembling real-life situations as well as situations relevant to economists. As behavior in these settings should only depend on risk attitudes, they will serve as benchmarks to contribute to the debate which risk elicitation methods are appropriate to predict behavior in these games. The benchmark games appeared in a randomized order. First, we used the same investment task as Charness and Gneezy (2010). Here, subjects could decide how much they wanted to invest in stocks and bonds out of an endowment of 10€. Subjects knew that any investment in bonds is a safe investment, and therefore they received the same amount they had invested in bonds as income. Additionally, the amount they invested in stocks was to be multiplied by 2.5 or lost completely with equal chance. Under EUT, this setting implies that both risk neutral and risk seeking decision makers should invest the entire amount. Thus, in order to be able to differentiate between them, we introduced another investment setting where the potential payment for stocks was 1.5 times the invested amount.

The third benchmark game was a first-price sealed-bid auction against a computerized opponent in line with Walker et al. (1987). Subjects could bid between 0.00€ and 20.00€ of their endowment, and they knew that the computer bid any amount between 0.00€ and 20.00€ with equal chance. The potential earnings (E_1 for subject 1) according to the bids (x_1; x_2) are:

$$E_1 = \begin{cases} 20 - x_1 & \text{if } x_1 > x_2 \\ 0 & \text{if } x_1 < x_2 \\ 20 - x_i \text{ or } E_1 = 0 \text{ (with 50\% chance)} & \text{if } x_1 = x_2 \end{cases}$$

Our benchmark games are deliberately chosen in such a way that risk is clearly relevant in the games, while being one step away from the artificial risk elicitation mechanisms. Therefore, all benchmark games are framed heavily, while still ensuring that risk attitudes should be the only factor driving a subject's decisions. The investment settings are very similar to the risk elicitation mechanisms described above in the sense that they resemble an SG method (with the difference that you choose your sure payoff and your lottery at the same time). The auction is more complex, as the optimal risk-neutral solution is harder to compute, but here you basically choose your own lottery, too. We therefore expect stronger correlation with the MPL methods for the investment games.

The experiment concluded with an extensive questionnaire. In order to incorporate survey-based measures, we asked subjects to provide an answer on a ten-point Likert-scale to the following two questions in line with Dohmen et al. (2011): "In general, are you a person who is fully prepared to take risks or do you try to avoid taking risks?" and "In financial situations, are you a person who is fully prepared to take risks or do you try to avoid taking risks?" The perceived complexity of these questions was elicited as well. In the questionnaire, we elicited the following socioeconomic factors: Age, gender, field of study, years of university education, nationality, high school grades in mathematics, monthly income and monthly expenditure. Furthermore, we elicited cognitive ability by conducting a cognitive reflection test (Frederick, 2005). Lastly, we assessed subjects' personalities in line with Rammstedt and John (2007), who provide a short measure of personality traits according to the BIG5[19] methodology introduced by Costa and McCrae (1992).

3 Results

We will first establish in sections 3.1 and 3.2 that the elicited risk parameter is highly dependent on the particular variant of MPL used because the overall distributions of switching points are very diverse and the rank correlations between the different methods are low in most circumstances. Section 3.3 analyzes the common features across methods. In section 3.4 we apply multiple measures to determine method quality. To this end, we first use benchmark games to let the data speak which risk elicitation methods predict behavior in these games best. In Section B, we will show which method produces the most stable results overall. Section 3.5 concludes with the result that the PGhigh method is the most stable method and that it has the highest predictive power.

[19] In the BIG5, personality is measured along five dimensions: Agreeableness, Conscientiousness, Extraversion, Neuroticism and Openness.

3.1 Overall Distributions are Different

According to EUT, a subject's behavior does not depend on which parameters are changed from row to row, as his underlying risk parameter value is constant. As the different versions of the MPL are calculated in such a way that the same switching point implies the same risk parameter interval, a consistent individual should have the same switching point in all versions of the MPL. This implies that the distributions of switching points should be the same across methods, barring some noise.

First, see Figure 1 for a graphical representation of the distributions. It is clearly visible at first glance that the distributions are not the same across all methods. For example, in the SGp method, most subjects would be classified as highly risk loving, whereas in the PGhigh method the majority of subjects would be classified as risk averse.
To verify whether distributions across methods are the same, we conduct two tests: a Friedman test, which shows that the means are not the same across methods ($p < 0.0001$), and a Kruskal-Wallis test, which shows that the distribution of answers is not the same across methods ($p < 0.0001$). We conclude that the switching points are, contrary to standard theory but in accordance with the literature, dependent upon the version of the particular MPL variation used.

To see which specific versions are significantly different from each other, we conduct a series of Wilcoxon tests, the natural pairwise analogue to the Kruskal-Wallis test. We use the Wilcoxon test to give a comparison of the distributions, as a difference in distributions is a more meaningful statistic here than a comparison of means. The p-values of the pairwise tests can be found in Table 4. Out of 55 pairwise comparisons, 28 comparisons indicate that methods are different at $p < 0.001$. Thirty-four (43) instances suggest that methods are different at $p < 0.01$ (0.05) significance levels.[20] To make sure that the differing results are not a product of fatigue or order effects, we also test whether CRRA-coefficients of methods that are encountered later in the experiment exhibit biases or more noise; the resulting tests show no significant order

[20] Note that one should be careful while reading this table and the ones following because of the presence of the multiple testing problem; therefore, we introduce a new notation in the tables: P-values lower than 0.001 are denoted by three stars, p-values lower than 0.01 are denoted by two stars and p-values lower than 0.05 are denoted by one star. $p < 0.001$ can be interpreted as significant, even when using the conservative Bonferroni correction (see Abdi, 2007).

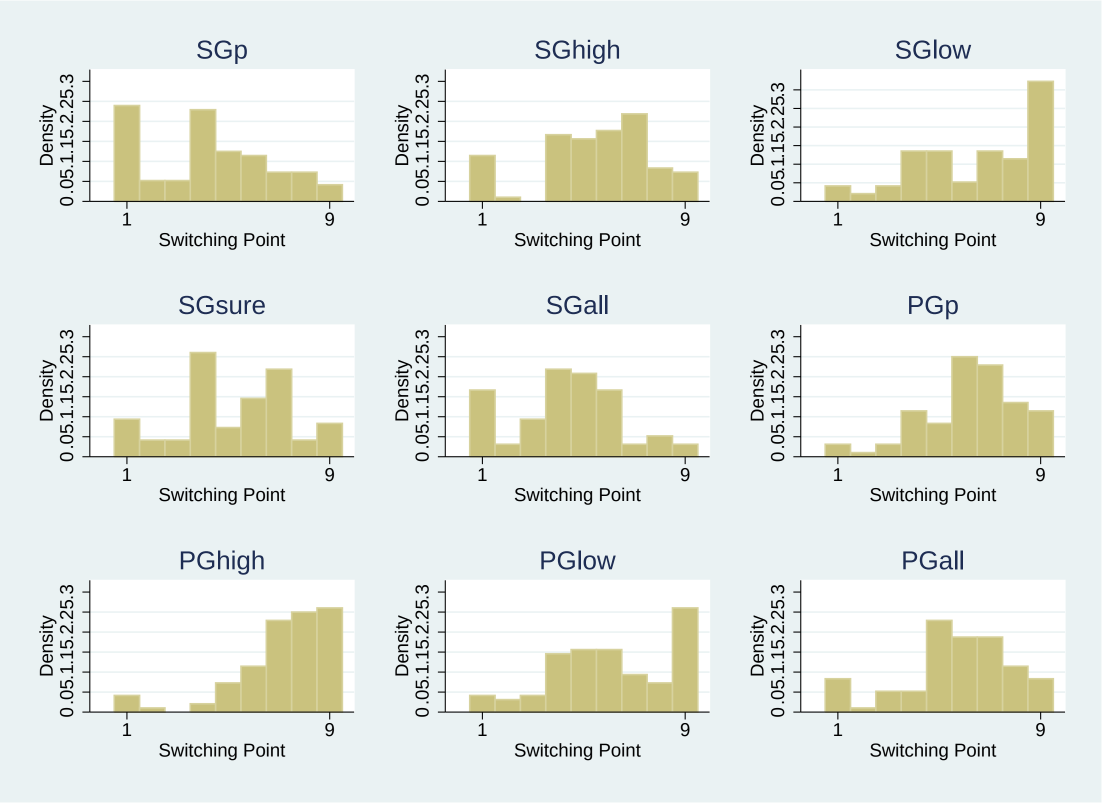


Figure 1 *Distributions of risk preferences; a low value indicates risk loving and a high value indicates risk averse behavior; x-axis: switching points (e.g. risk preferences) of subjects, where 1 means a subject switches from left to right in the first row and 9 means a subject never switches; y-axis: frequency of switching point*

effects overall and across methods.[21]

We conclude that different methods deliver significantly different results, and that the different versions of the MPL cannot be used interchangeably, as the estimated risk preference parameter depends heavily on the version used. Subjects can easily be classified as risk loving in one version and as risk averse in another. Of course we do not know a subject's true risk preferences, and therefore any of the methods might be able to classify a subject correctly. To provide an answer to this puzzle, see Section B, where we conduct a quality assessment of the different methods.

[21] See Table 11 in Appendix A.2.

Table 4: *Pairwise Wilcoxon test for equality of distribution*

	SGp	SGhigh	SGlow	SGsure	SGall	PGp	PGhigh	PGlow	PGall	GQ
SGhigh	.00***									
SGlow	.00***	.00***								
SGsure	.00***	.37	.00***							
SGall	.79	.00***	.00***	.01**						
PGp	.00***	.01**	.28	.00***	.00***					
PGhigh	.00***	.00***	.02*	.00***	.00***	.00***				
PGlow	.00***	.02*	.23	.01**	.00***	.68	.00***			
PGall	.00***	.31	.02*	.04*	.00***	.08	.00***	.39		
GQ	.02*	.03*	.00***	.29	.04*	.00***	.00***	.00***	.01**	
FQ	.00***	.64	.01**	.29	.00***	.02*	.00***	.04*	.36	.00***

*Notes: p-values of pairwise Wilcoxon tests are displayed; GQ: general question; FQ: financial question; stars are given as follows: *: p<0.05; **: p<0.01; ***: p<0.001*

3.2 Rank Correlations are Low

In this section we look at the rank correlation coefficients between the different methods and the questionnaire answers. If there are high rank correlations between the risk elicitation methods, one might argue that it is irrelevant which one is used if one intends to control for risk attitudes under any given circumstance. Rank correlations between the MPL methods and the questionnaire measures can be found in Table 5. We see that some of the correlations are significant, but only 11% of all pairwise comparisons in total if we test conservatively at $p < 0.001$ because of the multiple testing problem. Pay special attention to the fact that PGp, the most widely used method today, has no significant rank correlations with any of the other methods.[22] See also Table 10 in Appendix A.1 for standard correlations, which basically gives the same results as Table 5. These findings provide further evidence that the elicitation procedure should be chosen with care as the elicited risk aversion coefficient and also the relative ranking of subjects according to each method varies within broad boundaries.

[22] Also, the financial questionnaire (FQ) results have much higher correlations with the other methods than the general questionnaire (GQ) results, strengthening the argument that risk attitudes are domain specific.

Table 5: *Spearman rank correlation coefficients*

	SGp	SGhigh	SGlow	SGsure	SGall	PGp	PGhigh	PGlow	PGall	GQ
SGhigh	.46									
SGlow	.33***	.44***								
SGsure	.05	.22*	.26*							
SGall	.03	.18	-.03	.19						
PGp	.17	.15	.17	.21*	-.04					
PGhigh	.20	.39***	.21*	.03**	.21*	.25*				
PGlow	.31**	.28**	.25*	.19	-.02	.13	.21*			
PGall	.24	.21*	-.01	.08	.19	.04	-.01	.08		
GQ	.15	.13*	.06	-.12	.11	.02	.14	.04	.06	
FQ	.26*	.23*	.29*	.18	.10	-.04	.04	.24*	.13	.46***

*Notes: Table includes the nine different methods and the questionnaires (GQ: general questionnaire, FQ: financial questionnaire); stars are given as follows: *: p<0.05; **: p<0.01; ***: p<0.001*

3.3 Method Similarities

We have established in sections 3.1 and 3.2 that there are significant differences in the distributions of the risk elicitation methods. There are, however, some similarities that can be observed across methods: In Table 6, we classify MPLs according to whether the high payoff, the low payoff, the probabilities or the certainty equivalents change in the MPL table, whether the method has a certainty equivalent and whether the table was presented in a top-down or bottom-up format. Furthermore, we control for age, gender, cognitive reflection scores and the order in which the tables were presented. Column 1 shows the results for the first time a method was encountered, and column 2 for the repeated measurements.

We see that generally, methods that change the probabilities or methods that have a certainty equivalent classify subjects as more risk loving, while methods that change the low payoffs classify subjects as more risk-averse.[23] When a method is presented to subjects for the first time, changing the high payoff also classifies them as more risk-loving, while presenting

[23] In two MPL methods, PGall and SGall, multiple characteristics of the MPL table were changed at the same time. Consequently, in these methods the effects add up.

the table with ascending numbers seems to classify subjects as more risk-averse, although these two effects seem to vanish when presenting subjects with the same tables again. Note that we do not observe order effects, or significant effects of the control variables.

Table 6: *Similarities across all methods*

	Not Repeated	Repeated
High Payoff changes	$-.108^{**}$	$.052$
Low Payoff changes	$.208^{***}$	$.206^{***}$
Probability changes	$-.306^{***}$	$-.335^{***}$
Certainty Equivalent changes	$.086^{*}$	$-.039$
Has Certainty Equivalent	$-.496^{***}$	$-.504^{***}$
Top-Down Representation	$.085^{**}$	$.035$
Constant	$.835$	$.205$
R^2	$.121$	$.190$
Number of Observations	864	288

*Notes: OLS regressions clustered by individual subjects with one observation being the outcome from one answer of one subject in one method; dependent variable is the resulting CRRA-coefficient, with low scores indicating risk-loving behavior; independent variables on the left are dummies; nonsignificant controls for age, gender, order, BIG5 scores, income and CRT scores are included in the regressions but omitted in the table; first column gives results for the first time subjects encountered one of the nine methods, second column for the repeated measurements; stars are given as follows (differently than in the other tables, due to the absence of a multiple testing problem): *: p<0.10; **: p<0.05; ***: p<0.01*

3.4 Method Quality Indicators

We use two avenues to measure a method's quality: its predictive power (section 3.4.A) and its stability (section 3.4.B).

A Predictive Power

In order to see which method predicts behavior best in our benchmark games, we look at three statistics: the predictive power by simple OLS regression, the predictive power by Spearman rank correlation, and the absolute average deviation from the prediction.

In Table 7, we see the outcome of OLS regressions in the upper part, while controlling for personality measures and socioeconomic variables. In the lower part of Table 7 you see Spearman rank correlation coefficients, which we include because besides the absolute size of the elicited coefficients, the correct rank ordering of subjects is essential since these methods are often used to control for the role of risk attitude in various settings. The OLS regression can be understood as follows: The dependent variable is the outcome of a particular benchmark game, and the independent variables include the outcome in terms of the elicited risk aversion parameter ρ of one of the risk elicitation methods[24] plus all controls mentioned above.[25] The resulting coefficients in the investment games are negative because a higher ρ implies risk-averse behavior, and therefore lower investments and bids in the benchmark games; the reverse is true for the auction. The corresponding adjusted R^2 values can be found in parentheses below the coefficients.

The OLS regression equation is then given by

$$BG_{i,j} = \beta_0 + \beta_1 * MPL_j + \sum_{k=2}^{6} \beta_k * BIG5_k + \sum_{l=7}^{11} \beta_l * SE_l + \beta_{12} * CRT + \epsilon_i,$$

where i denotes the index of benchmark games (BG), j denotes the index of risk measures, MPL denotes the outcome of a risk elicitation method, *BIG5* denotes personality measures according to the *BIG5*, SE denotes the socioeconomic variables and CRT denotes the number of correct answers in the cognitive reflection test.

Additionally, we can calculate a point prediction in each of the benchmark games for each risk elicitation method (but not for the questionnaires). In Table 8 we report the absolute average deviations from these predictions, averaged over all three benchmark games according

[24] The results do not change qualitatively if we use the switching points as independent variables instead of ρ. This data is available upon request.

[25] It is not possible to add controls in the Spearman rank correlation.

to the formula

$$AAD = (\sum_{i=1}^{n} |H_i - H_i^*| + \sum_{i=1}^{n} |L_i - L_i^*| + \sum_{i=1}^{n} \frac{|A_i - A_i^*|}{2})/(3n),$$

where H_i denotes high investment game outcomes and H_i^* high investment game predictions (L stands for investment low and A for auction).[26]

Table 7: *Explanatory power*

	SGp	SGhigh	SGlow	SGsure	SGall	PGp	PGhigh	PGlow	PGall	GQ	FQ
	OLS coefficients										
Auction	.68	.52	.57	-.03	-.3	.03	0	.58	-.52	.13	-.08
	(.05)	(.03)	(.04)	(.02)	(.02)	(.02)	(.02)	(.04)	(.03)	(.03)	(.03)
Inv. Low	-.09	-.94*	-.48	.01	-.67	-.39	-1.48***	-.23	-.44	.3*	.09
	(.00)	(.03)	(.00)	(.00)	(.00)	(.00)	(.08)	(.00)	(.00)	(.02)	(.00)
Inv. High	-.9**	-.68	-.28	.22	.58	-.46	-.66	-.65	.2	.28**	.25**
	(.16)	(.12)	(.09)	(.09)	(.11)	(.09)	(.11)	(.12)	(.08)	(.13)	(.13)
	Spearman rank correlation coefficients										
Auct.	.23*	.09	.14	.17	.07	.11	.06	.16	-.13	.10	.11
Inv. Low	.02	.19	.17	.06	.06	.11	.36***	.12	.04	.19	.11
Inv. High	.28**	.28**	.05	.00	.03	.13	.26**	.23*	.09	.31**	.28**

*Notes: In the OLS regression, the dependent variable is the outcome in one of the four benchmark games, the independent variables are the outcome in terms of ρ from one method plus controls (age, gender, BIG5, CRT test, income, years of university education); the adjusted R^2 value for the regression can be found below a coefficient; Stars are given as follows: *: p<0.05; **: p<0.01; ***: p<0.001*

In the auction, none of the methods produce statistically significant results in the OLS regression. This is puzzling, as the auction can in itself be seen as a risk elicitation procedure, albeit with heavy framing. Recent literature, however, provides evidence that not only risk

[26] Note that we divide the deviation in the auction game by 2 because the choice range in the auction game is twice as high.

Table 8: *Deviations from predictions*

	SGp	SGhigh	SGlow	SGsure	SGall	PGp	PGhigh	PGlow	PGall
Deviation	1.91	2.41	2.19	2.03	2.11	2.27	**1.75**	2.17	2.11

Notes: Absolute average deviations from the predictions in the benchmark games.

attitudes but also other factors like regret aversion (Engelbrecht-Wiggans and Katok, 2008) could drive behavior in auctions. As far as the Spearman rank correlation is concerned, the SGp method is the only one that is rank correlated $(p < 0.05)$ with auction behavior.

In the investment games, the methods produce much better results. In the low investment setting, PGhigh has the biggest explanatory power, with SGhigh being a close follower. Note that it is surprising that PGhigh is the best predictor both in the regression and the rank correlation, as the investment games in themselves can be interpreted as standard gamble methods, so one would expect one of these methods to perform best.

In the high investment setting, many methods (PGhigh, PGlow, SGhigh, SGlow, SGp, and the questionnaires) are able to explain a part of the variance, with SGp being the one giving the best results $(p < 0.01)$. Note that in this setting, survey-based measures perform very well, so questionnaire measures seem to serve as good proxies for subjects' risk preferences in some circumstances. Note that the adjusted R^2 values are relatively low in general; we added the above mentioned controls to our regressions, which are not able to pick up much of the variation.[27]

As far as the deviations from the predictions are concerned, PGhigh performs best with an average deviation of 1.75 across all benchmark games with SGp and SGlow also having low deviations.[28]

In conclusion, PGhigh and SGp yield the best results in explaining behavior, with PGhigh

[27] In all regressions, none of the controls were significant at $p < 0.05$. This implies that behavior seems to primarily be driven by risk attitudes.

[28] Note that one might be concerned with this analysis because if a method generally classifies subjects as risk-averse, it is not surprising that it explains behavior well in the low investment setting, as subjects naturally behave risk-aversely in this setting due to the parameters. However, this critique is not valid for any method that provides good predictions across multiple benchmark games (e.g. PGhigh).

having the lowest deviation from the prediction of behavior in the benchmark games. We conclude that PGhigh has the highest predictive power with SGp being a close runner-up. Additionally, we relax our assumptions on CRRA and perform robustness checks taking CARA, DRRA, DARA, IRRA and IARA into account in Tables 23-34 in the Online Resource.[29] Furthermore, due to the ample evidence on the violations of EUT, we provide the same regressions by taking probability weighting[30], prospect theory[31] and cumulative prospect theory into consideration.[32] The results show that our findings remain quantitatively and qualitatively the same under different specifications. In general, we see similar explanatory power and in the vast majority of cases the same significance levels for the PGhigh and SGp methods, which confirm our findings. In some specifications, we even see that the coefficient for PGhigh becomes significant also for the investment games with low stakes, for example under DRRA. Nevertheless, a further justification is that some of the other methods (SGhigh, PGlow and questionnaire methods) lose significance (for example under PT, IRRA or IARA) under some of the above mentioned alternative theoretical foundations and functional forms.

B Stability Measures

In this section, we evaluate the stability of the different MPL representations. Remember that after our subjects had gone through all nine MPL methods, three of them were randomly chosen and presented to them again. A method can be described as stable if the given answers between the first and the second time a method was encountered are very similar. To analyze this similarity, we use three criteria: equality of overall distribution, equality of rank ordering and

[29] We used subjects' self-reported monthly income and expenditure as a proxy for their wealth. This was necessary, since our subject pool consisted of students who are not expected to have any wealth, but their monthly income or expenditure can serve as a good proxy for this purpose as they are expected to be highly correlated with their wealth (Persson and Tabellini, 1994) and social class.

[30] In the Online Resource, we report the regressions using the probability weighting function $w(p) = \frac{p^\gamma}{[p^\gamma + (1-p)^\gamma]^{\frac{1}{\gamma}}}$.

[31] As our lotteries are in the gains domain, prospect theory amounts to $u(c) = c^\alpha$.

[32] For prospect theory and cumulative prospect theory as well as probability weighting functions, we used the functional form and parameters provided in Tversky and Kahnemann (1992) and Quiggin (1982) to create the tables in the Online Resource; different specifications for PT provided in Camerer and Ho (1994) and in Wu and Gonzalez (1996), as well as for PW in Prelec (1998), do not change the results qualitatively. Specifications and parameters for DRRA and IRRA can be found in Andersen et al. (2012) and Saha (1993), respectively, for IARA and DARA in Saha (1993).

absolute average deviation between the first and second answers. For reasons of completeness, we also report the perceived complexity of each method.[33]

Table 9 reports these measures. In the first column we give p-values from a Kolmogorov-Smirnov test that evaluates whether the distributions of the first and the second time a method is encountered are the same. A significant p-value means that the distributions are significantly different from each other, indicating a low stability of overall distribution across a 30 minute time period. The second column gives the rank correlation between the first and second time a method was encountered. This measure is important because if a method's overall distribution merely shifted up or down without changing the rank ordering of subjects, this method can also be described as stable since the ordering of subjects remains the same. The third column reports the absolute average deviation (AAD) of subjects' answers when a particular method is presented to them again, compared to the first time – a lower value is therefore better. The last column gives the means of the perceived complexity of a method on a 1 to 10 Likert scale. To visualize these results, we also report the distributions of the differences in switching points between the first and the second time a method is encountered in Figure 2.

Any method that does not yield stable results over a 30 minute time period cannot be described as stable, and stability is a highly preferable characteristic in a risk measure. For the KS-test (column 1 in Table 9), stability relative to the other methods is indicated by a nonsignificant result: For any methods with a significant result, the overall distributions of answers are different between the first and the second time a method was encountered. Four of the methods have a nonsignificant p-value: SGp, SGhigh, PGp, PGhigh.

A significant rank correlation (column 2 in Table 9) also indicates a stable risk measure, indicating a shift in the distribution, but no change in rank ordering. We see that three of those four methods have significant rank correlations with $p < 0.01$: SGp, SGsure and PGhigh. A low absolute average deviation in answers is also an indicator of a stable risk measure, and the method with the lowest deviation is PGhigh, followed by PGp and SGsure. Concerning the complexity, we see that a method that is perceived as less complex does not necessarily imply more stability in answers, as SGlow has the lowest complexity rating, yet it is classified

[33] We do not use complexity as a stability measure, as the impact of a higher perceived complexity is not clear. On the one hand, one might argue that a higher measure in these categories implies noisier behavior, but on the other hand one might argue that a subject takes more time thinking about the problem at hand.

Table 9: *Stability Measures*

Method	KS-Test	Rank Corr.	AAD	Complexity
SGp	**.453**	.51***	1.60	3.42
SGsure	.003	.51***	**1.37**	3.92
SGhigh	**.644**	.39**	1.48	3.97
SGlow	.007	.35	1.96	3.20
SGall	.005	.16	1.8	4.81
PGp	**.240**	.23	**1.33**	4.21
PGhigh	**.879**	.45***	**1.24**	3.78
PGlow	.006	.25	2.04	4.29
PGall	.000	.19	1.85	5.75

*Notes: First column: P-values for a Kolmogorov-Smirnov test of equality of distributions; Second column: Rank correlation between the distributions of first and second answers (stars indicate significant rank correlation); Third column: Absolute average deviation (AAD) between the first and the second decision in the same method; Fourth column: Indicates a subject's perceived complexity of a method; Stars are given as follows: *: p<0.1; **: p<0.05; ***: p<0.01*

as unstable in all three categories. However, a general tendency of low complexity indicating more stability can be observed.

As far as a possible relationship between stability and the control variables (CRT and BIG5 scores, age, gender, income, years of university education) is concerned, no significant effects have been found, so the results will be omitted here.[34] Finally, we mention that the slight differences in the number of observations between the repetitions of particular methods – caused by the pseudo-random number generator – do not drive within-method consistency.

We conclude that PGhigh is the most stable method, as it is the only method that performs well in all three categories, with the overall distributions of switching points not significantly

[34] On the subject level, we tested whether the average deviation between the first and repeated decisions are related to any of the control variables, whether the standard deviation of CRRA scores across all methods is related to the control variables, and whether the average differences between the predicted and the actual outcomes in the benchmark games are related, both per method and overall. At most weakly significant results were found; the results are available from the authors upon request.

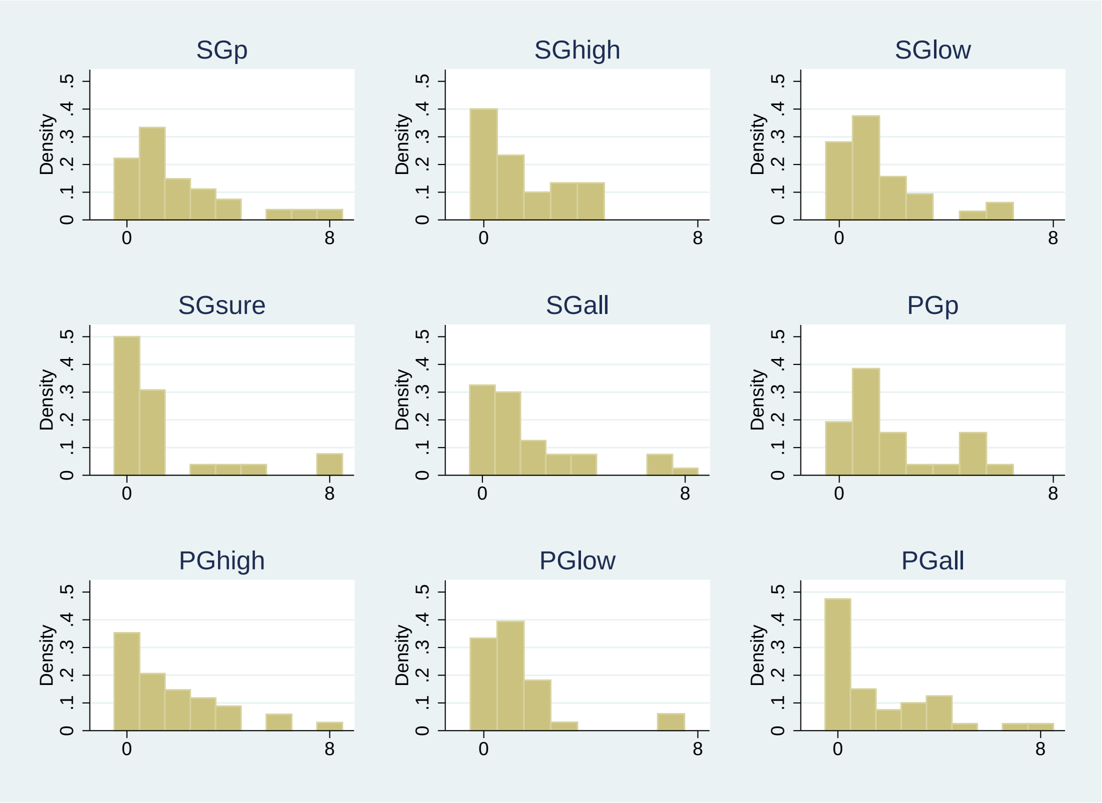


Figure 2 *Distributions of absolute differences in switching points between the first and the second time a method is encountered.*

different, high rank correlations and low average deviation. SGp, SGsure and PGp perform well in two of the three categories.

3.5 Results Conclusion

In the benchmark games, as far as predictive power is concerned, we conclude that PGhigh has the highest predictive power with SGp being a close second, irrespective of the assumed functional form or theoretical framework.[35] We conclude that only the PGhigh (Drichoutis and Lusk, 2012 & 2016), PGp (Holt and Laury, 2002), SGsure (Cohen et al., 1987; Abdellaoui et al., 2011) and SGp (Bruner, 2009) methods lead to consistent results within a 30-minute time frame, with the PGhigh method being by far the most consistent: The PGhigh method's

[35] For non-incentivized surveys, our data shows that eliciting preferences with general and financial questions is a relatively good predictor compared to several incentivized elicitation methods.

performance is superior to the other methods in terms of deviations from normative predictions, overall and relative stability across time, etc. Our findings are further supported by the fact that we controlled for personality traits, order effects, various socioeconomic factors and cognitive reflection in our analyses.

Therefore, we conclude that while SGp also has high predictive power and good stability in answers, the most stable MPL method with the highest predictive power is PGhigh, which corresponds to a method derived by Drichoutis and Lusk (2012, 2016) in our alternative interpretation.

4 Conclusion

We conducted a holistic assessment and analysis of MPL risk elicitation methods that are present in the economics literature with a sophisticated experimental design using a unified framework and representation method. Previous findings in the literature (Dave et al, 2010; Crosetto and Filippin, 2016; etc.) indicate that between-method consistency of particular methods is low. We confirm this finding by extending our analysis to all popular methods using the same representation. Furthermore, we show that distributional differences among methods are far from negligible. In addition, we investigate the time consistency of all these methods and document substantial differences in a 30 minute time period for most of the methods. All this implies that an arbitrary selection of a particular risk assessment method can lead to differing results and misleading revealed preferences. Thus, it matters which elicitation method is used by researchers in order to control for risk and other preferences.

Our main takeaway is that we provide a suggestion for which elicitation method to use based on objective criteria that assess within-method as well as between-method consistency and validity in real-world settings such as investments and auctions, and our suggestion is to use the PGhigh method by Drichoutis and Lusk (2012). This particular method performs best if we look at the absolute deviations from the normative predictions in benchmark games and also in terms of rank correlations. Furthermore, it yields highly correlated results within a 30-minute time frame in terms of individual deviations and overall distribution. These findings remain robust – in some cases even more pronounced – if we relax our assumptions on CRRA to alternative functions such as CARA, DRRA, DARA, IRRA and IARA. Moreover, our conclusions remain the same if we allow subjective probability weighting or if we estimate risk attitude parameters in line with prospect theory or cumulative prospect theory.

In a broader context, one should take care when choosing which risk elicitation method to use, especially if one aims to control for risk attitudes in potential real-world contexts such as investment into assets. To be taken into consideration are the nature of the task they intend to control for, trade-off effects between noise, exactness and simplicity. Moreover, we find that changing both the potential rewards and probabilities is perceived as relatively complex by subjects and yields inconsistent results. A further point to consider is that varying the potentially achievable minimum payoff seems to induce more risk-averse behavior while the presence of a certainty equivalent fosters risk taking. Cognitive ability, personality traits and other socioeconomic factors do not seem to be related to risk aversion nor to the extent of consistency we measured.

The debate between changing the probabilities or rewards (Bruner, 2009) seems to be far from settled as one of the methods in each context (PGhigh and SGp) delivers promising results. In addition, our findings might provide guidance in implementing other elicitation methods in the MPL format – e.g. loss aversion (Gächter et al., 2010), willingness to pay (Kahneman et al., 1990), individual discount rates (Harrison et al., 2002) – in terms of whether to vary probabilities, rewards or to use a certainty equivalent. On a final note we suggest that the relatively high variation in risk preferences across and within particular methods might not be mere artifacts – especially in light of other recent evidence (Andreoni et al., 2015). We encourage further research to shed light on the consistency of other preference elicitation mechanisms such as social preferences or overconfidence.

Acknowledgements

We thank the Vienna Center for Experimental Economics (VCEE), University of Vienna, for allowing us to run our experiments in their laboratory, and the Austrian Science Fund (FWF) under project S10307-G14 for their grateful support. We thank Jean-Robert Tyran, Wieland Müller, Erik Wengström, Karl Schlag, Owen Powell, James Tremewan, Rupert Sausgruber, Thomas Stephens and Stefan Minner for reading our paper, improving it with their comments and suggestions and guiding us in the right direction over the course of the project several times.

References

Abdellaoui, M., Driouchi, A. & L'Haridon, O. (2011). Risk aversion elicitation: Reconciling tractability and bias minimization. *Theory and Decision*, 71(1), 63-80.

Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons. In: N.J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* pp 103-107. Thousand Oaks, CA: Sage.

Andersen, S., Fountain, J., Harrison, G. W., Hole, A. R. & Rutström, E. E. (2012). Inferring beliefs as subjectively imprecise probabilities. *Theory and Decision*, 73(1), 161-184.

Andersen, S., Harrison, G. W., Lau, M. I. & Rutström, E. E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, 9(4), 383-405.

Andersen, S., Harrison, G. W., Lau, M. I. & Rutström, E. E. (2008a). Lost in state space: Are preferences stable? *International Economic Review*, 49(3), 1091-1112.

Andersen, S., Harrison, G. W., Lau, M. I. & Rutström, E. E. (2008b). Eliciting risk and time preferences. *Econometrica*, 76(3), 583-618.

Anderson, L. R. & Mellor, J. M. (2008). Predicting health behaviors with an experimental measure of risk preference. *Journal of Health Economics*, 27(5), 1260-1274.

Anderson, L. R. & Mellor, J. M. (2009). Are risk preferences stable? Comparing an experimental measure with a validated survey-based measure. *Journal of Risk and Uncertainty*, 39(2), 137-160.

Andersson, O., Holm, H. J., Tyran, J. R. & Wengström, E. (2016). Risk aversion relates to cognitive ability: Preferences or noise?, *Journal of the European Economic Association*, 14(5), 1129-54.

Andreoni, J., Kuhn, M. A. & Sprenger, C. (2015). Measuring time preferences: A comparison of experimental methods. *Journal of Economic Behavior and Organization*, 116(1), 451-464.

Attema, A. & Brouwer, W. (2013). In search of a preferred preference elicitation method: A test of the internal consistency of choice and matching tasks. *Journal of Economic Psychology*, 39(1), 126-140.

Beck, H. B. (1994). An experimental test of preferences for the distribution of income and individual risk aversion. *Eastern Economic Journal*, 20(2), 131-145.

Berg, J., Dickhaut, J. & McCabe, K. (2005). Risk preference instability across institutions: A dilemma. *Proceedings of the National Academy of Sciences of the United States of America*, 102(11), 4209-4214.

Binswanger, H. P. (1980). Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics*, 62(3), 395-407.

Blais, A. R. & Weber, E. U. (2006). A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1(1), 33-47.

Blanco, M., Engelmann, D., Koch, A. K. & Normann, H. (2010). Belief elicitation in experiments: is there a hedging problem?. *Experimental Economics*, 13(4), 412-438.

Bleichrodt, H. (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities, *Health Economics*, 11(5), 447-456.

Bleichrodt, H., Pinto, J. L. & Wakker, P. P. (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science*, 47(11), 1498-1514.

Bocqueho, G., Jacquet, F. & Reynaud, A. (2014). Expected utility or prospect theory maximisers? Assessing farmers' risk behaviour from field experiment data. *European Review of Agricultural Economics*, 41(1), 135-172.

Bruner, D. M. (2009). Changing the probability versus changing the reward. *Experimental Economics*, 12(4), 367-385.

Camerer, C. F., & Ho, T. H. (1994). Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty*, 8(2), 167-196.

Charness, G. & Gneezy, U. (2010). Portfolio choice and risk attitudes – An experiment. *Economic Inquiry*, 48(1), 133-146.

Charness, G., Gneezy, U. & Imas, A. (2013). Experiential methods: Eliciting risk preferences. *Journal of Economic Behavior and Organization*, 87(1), 43-51.

Charness, G. & Viceisza, A. (2016): Three risk-elicitation methods in the field: Evidence from rural Senegal, *Review of Behavioral Economics*, 3(2), 145-171.

Chiappori, P. & Paiella, M. (2011). Relative risk aversion is constant: Evidence from panel data. *Journal of the European Economic Association*, 9(6), 1021-1052.

Cohen, M., Jaffray, J.-Y. & Said, T. (1987). Experimental comparison of individual behavior under risk and under uncertainty for gains and for losses. *Organizational Behavior and Human Decision Processes*, 39(1), 1-22.

Cokely, E. T., Galestic, M., Schulz, E., Ghazal, S. & Garcia-Retamero, R. (2012). Measuring risk literacy – The Berlin Numeracy Test. *Judgment and Decision Making*, 7(1), 25-47.

Costa, P. T. & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Cox, J. C., Sadiraj, V. & Schmidt, U. (2015). Paradoxes and mechanisms for choices under risk. *Experimental Economics*, 18(2), 215-250.

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

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

Cubitt, R. P., Starmer, C. & Sugden, R. (1998). On the validity of the random lottery incentive system. *Experimental Economics*, 1(2), 115-131.

Dave, C., Eckel, C. C., Johnson, C. A. & Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, 41(3), 219-243.

De Véricourt, F., Jain, K., Bearden, J. N. & Filipowicz, A. (2013). Sex, risk and the newsvendor. *Journal of Operations Management*, 31(1-2), 86-92.

Dohmen, T., Falk, A., Huffman, D. & Sunde, U. (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review*, 100(3), 1238-1260.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J. & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants and behavioral consequences. *Journal of the European Economic Association*, 9(3), 522-550.

Drichoutis, A. & Lusk, J. (2012). Risk preference elicitation without the confounding effect of probability weighting. Working paper. Munich Personal RePEc Archive, Munich, Germany.

Drichoutis, A. & Lusk, J. (2016). What can multiple price lists really tell us about risk preferences? *Journal of Risk and Uncertainty*, 53(2), in press.

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

Eckel, C. C. & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, 23(4), 281-295.

Eckel, C. C. & Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior and Organization*, 68(1), 1-17.

Engelbrecht-Wiggans, R. & Katok, E. (2008). Regret and feedback information in first price sealed-bid auctions. *Management Science*, 54(4), 808-819.

Farquhar, P. H. (1984). State of the art – Utility assessment methods. *Management Science*, 30(11), 1283-1300.

Fellner, G. & Maciejovsky, B. (2007). Risk attitude and market behavior: Evidence from experimental asset markets. *Journal of Economic Psychology*, 28(3), 338-350.

Fischbacher, U. (2007): z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics*, 10(2), 171-178.

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

Goeree, J. K., Holt, C. A. & Palfrey, T. R. (2003). Risk averse behavior in generalized matching pennies games. *Games and Economic Behavior*, 45(1), 97-113.

Gonzalez, R. & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, 38(1), 129-166.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1), 114-125.

Harrison, G. W., Humphrey, S. J. & Verschoor, A. (2010). Choice under Uncertainty: Evidence from Ethiopia, India and Uganda. *The Economic Journal*, 120(543), 80-104.

Harrison, G. W., Johnson, E., McInnes, M. M. & Rutström, E. E. (2005). Temporal stability of estimates of risk aversion. *Applied Financial Economics Letters*, 1(1), 31-35.

Harrison, G. W., Lau, M. I. & Rutström, E. E. (2009). Risk attitudes, randomization to treatment, and self-selection to experiments. *Journal of Economic Behavior and Organization*, 70(3), 498-507.

Harrison, G. W., Lau, M. I. & Williams, M. B. (2002). Estimating individual discount rates in Denmark: A field experiment. *American Economic Review*, 92(5), 1606-1617.

Harrison, G. W., List, J. A. & Towe, C. (2007). Naturally occurring preferences and exogenous laboratory experiments: A case study of risk aversion. *Econometrica*, 75(2), 433-458.

Harrison, G. W. & Rutström, E. E. (2008). Risk aversion in the laboratory. In: J. C. Cox and G. W. Harrison (Eds.), *Risk Aversion in Experiments*, (pp. 41-196). Research in Experimental Economics 12. Bingley, UK: Emerald.

Harrison, G. W. & Rutström, E. E. (2009). Expected utility theory and prospect theory: one wedding and a decent funeral. *Experimental Economics*, 12(2), 133-158.

Hershey, J. C., Kunreuther, H. C. & Schoemaker, P. J. H. (1982). Sources of bias in assessment procedures for utility functions. *Management Science*, 28(8), 936-954.

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

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

Holt, A. C. & Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. *American Economic Review*, 95(3), 902-912.

Isaac, R. M. & James, D. (2000). Just who are you calling risk averse? *Journal of Risk and Uncertainty*, 20(2), 177-187.

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

Kahneman, D., Knetsch, J. L. & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, 98(6), 1325-1348.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263-291.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R. & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking – BART. *Journal of Experimental Psychology: Applied*, 8(2), 75-84.

Levy, H. (1994). Absolute and relative risk aversion: An experimental study. *Journal of Risk and Uncertainty*, 8(3), 289-307.

Lönnqvist, J-E., Verkasalo, M. J., Walkowitz, G. & Wichardt, P. C. (2015). Measuring individual risk attitudes in the lab: Task or ask? An empirical comparison. *Journal of Economic Behavior and Organization*, 119(1), 254-266.

Lusk, J. L. & Coble, K. H. (2005). Risk perceptions, Risk preference, and acceptance of risky food. *American Journal of Agricultural Economics*, 87(2), 393-405.

Mador, G., Sonsino, D. & Benzion, U. (2000). On complexity and lotteries' evaluation – three experimental observations. *Journal of Economic Psychology*, 21(6), 625-637.

Murnighan, J. K., Roth, A. E. & Schoumaker, F. (1988). Risk aversion in bargaining: An experimental study. *Journal of Risk and Uncertainty*, 1(1), 101-124.

Persson, T., & Tabellini, G. (1994). Is inequality harmful for growth?. *American Economic Review*, 84(3), 600-621.

Poulton, E. C. (1989). *Bias in quantifying judgments*. Hove, UK: Erlbaum.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, 32(1-2), 122-136.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66(3), 497-527.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, 3(4), 323-343.

Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the BIG Five Inventory in English and German. *Journal of Research in Personality*, 41(1), 203-212.

Reynaud, A. & Couture, S. (2012). Stability of risk preference measures: Results from a field experiment on French farmers. *Theory and Decision*, 73(2), 203-221.

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

Saha, A. (1993). Expo-power utility: A 'flexible' form for absolute and relative risk aversion. *American Journal of Agricultural Economics*, 75(4), 905-913.

Tanaka, T., Camerer, C. F. & Nguyen, Q. (2010). Risk and time preferences: Linking experimental and household survey data from Vietnam. *American Economic Review*, 100(1), 557-571.

Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297-323.

Visschers, V. H. M., Meertens, R. M., Passchier, W. W. F. & De Vries, N. N. K. (2009). Probability information in risk communication: A review of the research literature. *Risk Analysis*, 29(2), 267-287.

Von Gaudecker, H. M., van Soest, A. & Wengström, E. (2011). Heterogeneity in risky choice behavior in a broad population. *American Economic Review*, 101(2), 664-694.

Wakker, P. P. (2008). Explaining the characteristics of the power (CRRA) utility family. *Health Economics*, 17(12), 1329-1344.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge, UK: Cambridge University Press.

Wakker, P. P., & Deneffe, D. (1996). Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science*, 42(8), 1131-1150.

Walker, J. M., Smith, V. L. & Cox, J. C. (1987). Bidding behavior in first-price sealed bid auctions: Use of computerized Nash competitors. *Economics Letters* 23(3), 239-244.

Weber, E. U., Blais, A-R. & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15(4), 263-290.

Wilkinson, L. & Wills G. (2005). *The grammar of graphics*. Berlin, Germany: Springer

Wu, G. & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management Science*, 42(12), 1676-1690.

A Appendix

A.1 Robustness Checks

Table 10 shows standard correlation coefficients between methods; the results are qualitatively the same as in Table 5.

Table 11 tests for linear order effects in the data: As the nine methods were presented to subjects in a randomized order, risk aversion or noise might increase or decrease as a function of previously seen MPLs. Coefficients in Table 11 can be interpreted as a change in standard deviation (i.e. noise) of CRRA coefficients or changes in CRRA coefficients themselves as a function of order, i.e. the number of previously encountered MPLs. No significant effects were found.

For robustness checks on functional form (PT, CARA, DRRA, IRRA, IARA, DARA and probability weighting), please refer to the Online Resource, Tables 23-34.

Table 10: *Correlation coefficients between the methods*

	SGp	SGhigh	SGlow	SGsure	SGall	PGp	PGhigh	PGlow	PGall	GQ
SGhigh	.46***									
SGlow	.37***	.46***								
SGsure	.01	.18	.25*							
SGall	.02	.12	0	.16						
PGp	.13	.12	.15	.23**	-.08					
PGhigh	.07	.27**	.10	.26*	.12	.26*				
PGlow	.27**	.21*	.20	.17	-.04	.12	.10			
PGall	.17	.16	-.06	.04	.17	-.01	-.08	.03		
GQ	.12	.16	0	-.14	.09	-.07	.12	.02	.01	
FQ	.25**	.17	.27**	.19	.07	-.05	-.02	.20	.03	.46**

*Notes: This table shows standard correlations as opposed to Spearman rank correlations as in Table 5. SG stands for "standard gamble" and PG for "paired gamble". Our conclusions remain qualitatively the same. Stars are given as follows: *: p<0.05; **: p<0.01; ***: p<0.001*

Table 11: *Testing for order effects – No significant effects*

	Dependent Variable	
Method	CRRA Standard Deviation	CRRA
PGhigh	-0.001	0.006
PGlow	-0.013	0.019
PGp	-0.021	0.028
PGall	-0.013	0.013
SGhigh	-0.007	-0.014
SGlow	-0.023	0.029
SGsure	0.031	-0.037
SGp	0.021	-0.020
SGall	-0.011	-0.004
Overall	-0.004	0.000

*Notes: Reported coefficients for OLS regressions; the dependent variable is the standard deviation (column 1) of the CRRA-coefficient or the CRRA score (column 2) across all subjects in one particular order, the independent variable is the order in which a method appeared, i.e. the number of previously encountered MPLs; stars are given as follows: *: p<0.1; **: p<0.05; ***: p<0.01*

A.2 Comparison of our Results to the Results in Previous Studies

In Table 12 we see the differences in the mean values of CRRA risk coefficients between previous studies and our results for each method, where we find that several studies deliver significantly different results to ours. This is not surprising for two reasons: First, risk elicitation methods are very noisy in general. For example the same method with the same representation delivers significantly different results in Crosetto and Filippin (2013) and Crosetto and Filippin (2016), or the task by Holt and Laury (2002), which delivered highly heterogeneous results in past studies as Table 12 shows. Second, framing and representation are vastly different in most studies when compared to our study. Furthermore, in the studies by Eckel and Grossman (2008), Reynaud and Couture (2012) and Crosetto and Filippin (2016) the risk loving domain is not covered in the SGall task; that and the pull-to-center effect drives the risk estimates to be higher, which is also suggested by Bleichrodt (2002) and Andersen et al. (2006).

Table 12: *Comparison of results to previous studies*

	Our study		Previous Studies				t-test
Method	Mean	SD	Mean	SD	Subjects	Study	p-value
PGhigh	0.87	0.56	0.35	0.18	100	Drichoutis and Lusk (2012)	.001
PGlow	0.57	0.67	0.35	0.18	100	Drichoutis and Lusk (2012)	.002
PGp	0.62	0.54	0.32	0.41	175	Holt and Laury (2002)	.001
			0.23	0.14	39	Abdellaoui et al. (2011)	.001
			0.59	0.07	100	Drichoutis and Lusk (2012)	.145
			0.39	0.54	78	Dulleck et al. (2015)	.006
			0.43	0.6	444	Crosetto and Filippin (2016)	.004
			0.62	0.8	268	Andersen et al. (2008b)	1
			0.67	0.57	881	Dave et al. (2010)	.455
PGall	0.47	0.62	0.82	-	86	Lejuez et al. (2002)	-
			1.13	0.64	444	Crosetto and Filippin (2016)	.001
			0.7	0.83	444	Crosetto and Filippin (2013)	.011
SGhigh	0.4	0.65	0.51	0.59	157	Bruner (2009)	.168
SGlow	0.69	0.68					
SGsure	0.31	0.66	0.2	0.08	39	Abdellaoui et al. (2011)	.302
SGp	0.02	0.72	0.45	0.45	157	Bruner (2009)	.001
SGall	0.07	0.63	0.6	0.59	256	Eckel and Grossman (2008)	.001
			0.73	0.9	30	Reynaud and Couture (2012)	.001
			0.694	0.33	444	Crosetto and Filippin (2016)	.001

Notes: mean and standard deviation in terms of CRRA-coefficients; $N = 96$ in our study; PGp by Crosetto and Filippin (2016) follows the method in Lejuez et al. (2002); in Lejuez et al. (2002) no standard deviation was reported

The Sequential Hotelling Game - Slowly learning the Equilibrium Outcome

Alexander Rabas[*]

June 23, 2017

Abstract

This article presents a theoretical and experimental investigation of a 3-player sequential-entry variant of Hotelling's locational choice model (1929). Martin Osborne and Amoz Kats offer a conjecture about the unique subgame-perfect Nash equilibrium (SPNE) outcome in this game, which I prove for $n = 3$: The first and the last player enter at the median, and the middle player opts out. When used to model political elections, the character of this equilibrium is then related to Duverer's Law, as a two-party system will emerge. Testing this conjecture in the lab reveals that in the beginning, the first and middle player keep out the last player. However, after many repetitions, play converges toward the unique SPNE outcome.

Keywords: Hotelling, Experiment, Sequential, Duverger's Law

JEL Classification: C72 · C91

[*] University of Vienna, Doctoral School of Economics Vienna, Department of Economics
alexander.rabas@gmail.com

1 Introduction

In this paper I analyze a game that was first presented by Martin Osborne and Amoz Kats, which I will henceforth call the sequential Hotelling game:[1]

> "Each player $1, \dots, n$ chooses a member of the set $[0,1] \cup OUT$ (i.e. either chooses a "location" or opts out). The choices are made sequentially (starting with player 1), and every player is perfectly informed at all times. The outcome of the game is determined as follows. After all players have chosen their actions, each player who has chosen a location receives votes from a continuum of citizens; the player who receives the most votes wins. The distribution of citizens' ideal points is nonatomic, with support [0,1]. A player who chooses the same position x as $k-1$ other players obtains the fraction $1/k$ of the votes of all the citizens whose ideal points are closer to x than to any other chosen location. [...] Each player obtains the payoff 0 if she chooses OUT, the payoff $1/k$ if she is among the k players who receive the maximal fraction of votes, and -1 otherwise."[2,3]

Martin Osborne and Amoz Kats offer a conjecture[3,4] about the subgame-perfect Nash-equilibrium (SPNE) outcome in this sequential Hotelling game for an arbitrary number of players n:

Osborne-Kats Conjecture. The sequential Hotelling game has a unique SPNE outcome, in which players 1 and n choose the median location m and all other players choose OUT.

This game is interesting in two ways: First, as the only game (to my knowledge) to feature a first-mover and a last-mover advantage simultaneously, it is unique in a game theoretical sense. Second, the specific character of the equilibrium is related to Duvergers Law:[5] When the game is interpreted as modeling the location decisions of political actors on a left-right spectrum, a

[1] The sequential Hotelling game is a variant of Hotelling's locational choice model (1929) and its refinement by Duverger (1954).

[2] In the original game description the term "position" is used instead of the term "location"; this was changed for consistency reasons because I use the term location in the experiment.

[3] Freely available on Martin Osborne's homepage under http://www.economics.utoronto.ca/osborne/research/CONJECT.HTM

[4] A proof for the general case does not exist yet.

[5] A plurality voting system often leads to a two-party system; Duverger (1954).

two-party system will emerge.

I analyze the three-player variant of this game, for which I derive all SPNE and prove the conjecture for $n = 3$.[6] Furthermore, I implement the game under laboratory conditions and conduct an experiment, in which the Osborne-Kats conjecture makes a strong behavioral prediction in the lab, as the SPNE outcome is unique.[7] Subjects play the sequential Hotelling game as a finitely repeated game, with the treatments differing (mostly) in game length.

Initial play in the experiment leads to an outcome that favors players 1 and 2, which is not in accordance with the SPNE in the sequential Hotelling game. As the game progresses, however, players learn to best respond, resulting in convergence toward the unique SPNE outcome after many repetitions of the game.

As Hotelling's original locational choice model (1929) and its derivations are highly relevant in the fields of political science and industrial organization, numerous theory papers exist on this topic; see Osborne (1995) for a general review. While the simultaneous move case and its variants have also been analyzed experimentally,[8] changes in timing, i.e. the sequential entry case, have only been analyzed theoretically,[9] so never empirically or experimentally. Therefore, my contribution to the literature is a theoretical and experimental investigation of the Hotelling game with sequential entry.

[6] Detailed arguments for the unique SPNE outcome in the special case of $n = 3$ were already made in Osborne (2004) and on http://www.economics.utoronto.ca/osborne/research/ARG.HTM, but to the author's knowledge a formal proof does not exist in the literature.

[7] The lab implementation is not exactly the same as the sequential Hotelling game, as the voter support is continuous in theory but must necessarily be discrete in the lab; more on that in Section 2.2.

[8] See Brown-Kruse, Cronshaw and Schenk (1993), Brown-Kruse and Schenk (2000), Collins and Sherstyuk (2000), Huck, Müller and Vriend (2000), Barreda-Tarrazona et al. (2011) and Kephart and Friedman (2015).

[9] See Prescott and Visscher (1977), Neven (1987), Eiselt and Laporte (1997), Osborne (2004), Rabas (2011), Bandyopadhyay et al. (2016), as well as Kress and Pesch (2012) for an overview.

2 Theory

2.1 The Sequential Hotelling Game

Each of the players $i = 1, \ldots, n$ chooses as his action a_i an element of the set $[0,1] \cup \{OUT\}$. That is, each player either chooses a location ($a_i \in [0,1] \setminus OUT$) or opts out ($a_i = OUT$). The choices are made sequentially starting with player 1 and ending with player n, and every player observes all previous choices.

The outcome of the game is then determined as follows: After all players have chosen their actions, each player i who has chosen a location $a_i \neq OUT$ receives vote shares $v_i(a_1, a_2, a_3) \in (0,1]$, where $\sum_{i=1}^{n} v_i = 1$, from a continuum of voters. The player(s) who receive(s) the most vote shares $v^{max} = max(v_1, v_2, \ldots, v_n)$ win(s); a player who has chosen $a_i = OUT$ receives no votes and cannot win. Each voter simply votes for the player whose chosen location $a_i \neq OUT$ is closest to the voter's ideal location, and the distribution of voters' ideal locations is uniform along the interval $[0,1]$ and nonatomic.[10] Furthermore, if a player i has chosen the same location a_i as $z-1$ other players, he obtains the fraction $v_i = 1/z$ of the votes of all voters whose ideal location is closer to a_i than to any other chosen location.

Each player i then obtains payoff π_i according to the following formula, where s denotes the number of players with $v_i = v^{max}$, i.e. $s = |\{i \in \{1, \ldots, n\} \, | v_i = v^{max}\}|$:

$$\pi_i = \begin{cases} 0 & \text{if } a_i = OUT \\ 1/s & \text{if } a_i \in [0,1] \text{ and } v_i = v^{max} \\ -1 & \text{if } a_i \in [0,1] \text{ and } v_i < v_i^{max}. \end{cases}$$

That is, each player obtains payoff 0 if he chooses $a_i = OUT$, payoff $1/s$ if he is among the s players who receive the maximal share of votes, and -1 if there exists a player who has more votes. This implies that each player wants to enter the competition if and only if he has some chance of winning.

For readability, I also introduce two definitions:

[10] This means that if a voter's favorite location is x^*, he is indifferent between the locations $x^* - s$ and $x^* + s$. This also means that the voters do not vote strategically, and voting is sincere.

Definition 1. I define "choosing a location" and "entering the game" as choosing an $a_i \neq OUT$. Furthermore, in the sequential Hotelling game, I define choosing an $a_i < t$ as $a_i \in [0, t)$. Finally, if $a_i = OUT$, player i "stays out of the game".

Definition 2. In the Sequential Hotelling Game, I call a player "winning" if he has payoff $\pi_i > 0$, and I call a player "losing" if he has payoff $\pi_i = -1$. Furthermore, a player "wins alone" if he has strictly more votes than any other player.

A SPNE for the Sequential Hotelling Game

In this paper, I will look at the case of $n = 3$. The different subgame-perfect Nash-equilibria for $n = 3$ are characterized as follows:[11]

$$a_1^* = 0.5,\ a_2^*(a_1) = \begin{cases} 0.5 & \text{if } a_1 = OUT \\ [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1] & \text{if } a_1 < \frac{1}{6} \\ [\frac{2-a_1}{3}, 1 - a_1) & \text{if } \frac{1}{6} \leq a_1 \leq 0.5 \\ OUT & \text{if } a_1 = 0.5, \end{cases} \tag{1}$$

while player 3 chooses according to the following rules:

1. If the set $A = \{a_3 | v_3 > max(v_1, v_2)\}$ is nonempty, i.e. if player 3 can attain $v_3 > max(v_1, v_2)$ by choosing some $a_3 \in [0, 1]$, he chooses one of these payoff-maximizing choices.

2. If set A is empty and the set $B = \{a_3 | v_3 = max(v_1, v_2)\}$ is nonempty, i.e. player 3 can attain $v_3 = max(v_1, v_2)$ by choosing some $a_3 \in [0, 1]$, he chooses one of them.

3. If both sets A and B are empty, $a_3 = OUT$.

[11] Because of the symmetrical nature of the game around the median location 0.5, there are certain symmetries in this game. In general, if we make any statement concerning outcome, vote shares or best responses about a choice triple (a_1, a_2, a_3), the same statement is still true if we consider the choice triple $(1 - a_1, 1 - a_2, 1 - a_3)$ (here I define for $a_i = OUT$ that $1 - a_i = OUT$). Or in other words, $\pi_i(a_1, a_2, a_3) = \pi_i(1 - a_1, 1 - a_2, 1 - a_3)$ and $v_i(a_1, a_2, a_3) = v_i(1 - a_1, 1 - a_2, 1 - a_3)\ \forall\ i$. For best responses, it holds that if a_2^* is a best response to a_1 (given a_3^*), then $1 - a_2^*$ is a best response to $1 - a_1$ (given $1 - a_3^*$). For player 3, if a_3^* is a best response to (a_1, a_2), $1 - a_3^*$ is a best response to $(1 - a_1, 1 - a_2)$. Therefore, cases of $a_1 = 0.5 + s$ are symmetrical to $a_1 = 0.5 - s$ (for $s \leq 0.5$) , and I can omit all cases $a_1 > 0.5$ w.l.o.g.

It is important to note that while the subgame-perfect Nash-equilibrium outcome is unique, the equilibria themselves are not.[12] In equation (1), we can clearly see why this is the case: After many histories, the best responses by players 2 and 3 are not unique off the equilibrium path. We see for example that for player 2, there is an infinite number of best responses given $a_1 < \frac{1}{6}$. The same is true for player 3, as there are many cases where there is a range of best responses following (a_1, a_2), giving an infinite number of subgame-perfect Nash equilibria.

The important fact is that according to *any* SPNE, play along the equilibrium path consists of $a_1 = 0.5$, $a_2 = OUT$, $a_3 = 0.5$, in accordance with the Osborne-Kats-conjecture. The derivation of the above SPNE as well as the uniqueness of the outcome can be found in appendix A.1.

2.2 Lab Implementation

The problem with the transition of the sequential Hotelling game as described above to a laboratory environment is that a player's action space is continuous in theory, but necessarily discrete in the lab. Furthermore, it is crucial that player 3 has the option to choose an a_3 closer to the median than player 2.[13]

I solve this problem by giving subjects different discrete action spaces depending on their positions, i.e. the order in which the players choose their actions $a_i \in X_i$:

- A player in position 1 can choose from $X_1 = \{OUT, 1, 9, 17, 25, 33, 41, 49\}$
- A player in position 2 can choose from $X_2 = \{OUT, 1, 5, 9, \dots, 41, 45, 49\}$
- A player in position 3 can choose from $X_3 = \{OUT, 1, 3, 5, \dots, 45, 47, 49\}$

In this way, subjects who choose later have more options than subjects who choose earlier. The set $X = \{1, 2, 3, \dots, 47, 48, 49\}$ contains all locations in the game. The locations available to

[12] This fact is also pointed out in Osborne (2004).

[13] If this were not the case, the unique SPNE outcome for $n = 3$ would be given by $a_1^* = OUT$, $a_2^* = 0.5$, $a_3^* = 0.5$, a different equilibrium outcome. The reasoning for this is the following: Assuming player 1 chooses $a_1 = 0.5$, player 2 can adopt the location $a_2 = 0.5 - \epsilon$ closest to 0.5 that is possible, while player 3 now has no option to locate closer to the middle than player 3, so he chooses $a_3 = 0.5 + \epsilon$ and shares the win with player 2. Therefore, player 1 will stay out of the game, and the other two players locate at the median. A formal proof is available from the author upon request.

be chosen by each player i are defined as $X_i \backslash \{OUT\}$.

The outcome of the game is determined similarly as in the sequential Hotelling game: After all players have chosen their actions, each player who has chosen a location, i.e. $a_i \in X_i \backslash \{OUT\}$, receives a number of points $v_i(a_1, a_2, a_3) \in (0, 48]$, while a player who chooses $a_i = OUT$ receives no points, i.e. $v_i = 0$. Each location $x \in [2, 3, 4, \dots, 46, 47, 48]$ is worth one point, and the locations on the edges ($x = 1$ and $x = 49$) are worth half a point. Therefore, the sum of points to be gained is 48. Each player receives points from each location x that is closer to his chosen location a_i than to any other location that was chosen by another player. If a player i has chosen the same location a_i as $z - 1$ other players, he obtains the fraction $v_i = 1/z$ of the points from locations that are closer to a_i than to any other chosen location. Furthermore, if an unchosen location is equally distant between two chosen locations a_i and a_j, the point for this unchosen location is split evenly between the players who have chosen a_i and a_j. The player(s) who receive(s) the largest share of points $v^{max} = max(v_1, v_2, v_3)$ win(s), given that $v^{max} \neq 0$.

Each player i then obtains payoff π_i according to the following formula, where s denotes the number of players with $v_i = v^{max}$ who choose a location, i.e. $s = |\{i \in \{1, 2, 3\} \ |v_i = v^{max} \cap v_i \neq 0\}|$:

$$\pi_i = \begin{cases} 0.25 & \text{if } v_i = 0 \\ 2/s & \text{if } v_i = v^{max} \text{ and } v_i \neq 0 \\ 0.05 & \text{if } v_i < v^{max} \text{ and } v_i \neq 0. \end{cases}$$

That is, each player gets 0.25 if he chooses $a_i = OUT$, payoff $2/j$ if the player is among the s players who receive the maximal share of points, and 0.05 if there exists a player who has more points.[14]

The definitions of winning, losing and entering the game for the lab game are as follows:

Definition 3. I define choosing an $a_i < t$ as choosing an $a_i \in \{X_i \cap [1, t)\}$ in the lab game.[15]

[14] For the lab game to have qualitatively the same incentives as the sequentlial Hotelling game, two conditions must be fulfilled as far as the payoff parameters are concerned: If player i wins, his payoff must always be higher than if he chooses $a_i = OUT$ (which is satisfied here, as $2/n$ is always higher than 0.25), and if a player chooses OUT he must have a higher payoff than if he loses, which is also satisfied.

[15] For example, if player 1 chooses an $a_1 < 17$, he either chooses $a_1 = 1$ or $a_1 = 9$.

Definition 4. In the lab game, I call a player "winning" if he has payoff $\pi_i > 0.25$, and I call a player "losing" if he has payoff $\pi_i = 0.05$. Furthermore, a player "wins alone" if he has strictly more votes than any other player.

Definition 5. In the lab game, I say that a player "enters the game" if that player chooses any location, i.e. $a_i \in X_i \setminus OUT$, and a player "stays out of the game" if $a_i = OUT$.

A SPNE in the Lab Implementation

Analogous to the the sequential Hotelling game, the SPNE for the lab implementation is not unique, but the SPNE outcome is. The SPNE for the lab implementation is characterized by[16]

$$a_1^* = 25, \; a_2^*(a_1) = \begin{cases} 25 & \text{if } a_1 = OUT \\ 37 & \text{if } a_1 = 1 \\ \{33, 37\} & \text{if } a_1 = 9 \\ 29 & \text{if } a_1 = 17 \\ OUT & \text{if } a_1 = 25, \end{cases}$$

while player 3 chooses according to the following rule:

1. If the set $A = \{a_3 | v_3 > max(v_1, v_2)\}$ is nonempty, i.e. if player 3 can attain $v_3 > max(v_1, v_2)$ by choosing some $a_3 \in X_3$, he chooses one of these payoff-maximizing choices.

2. If set A is empty and the set $B = \{a_3 | v_3 = max(v_1, v_2)\}$ is nonempty, i.e. player 3 can attain $v_3 = max(v_1, v_2)$ by choosing some $a_3 \in X_3$, he chooses one of them.

[16] Similar to footnote 11, because of the symmetrical nature of the lab game around the median location 25, there are certain symmetries in this game. In general, if we make any statement concerning outcome, vote shares or best responses about a choice triple (a_1, a_2, a_3), the same statement is still true if we consider the choice triple $(50 - a_1, 50 - a_2, 50 - a_3)$ (here I define for $a_i = OUT$ that $50 - a_i = OUT$). Or in other words, $\pi_i(a_1, a_2, a_3) = \pi_i(50 - a_1, 50 - a_2, 50 - a_3)$ and $v_i(a_1, a_2, a_3) = v_i(50 - a_1, 50 - a_2, 50 - a_3) \; \forall \; i$. For best responses, if a_2^* is a best response to a_1 (given a_3^*), then $1 - a_2^*$ is a best response to $1 - a_1$ (given $1 - a_3^*$). For player 3, if a_3^* is a best response to (a_1, a_2), $50 - a_3^*$ is a best response to $(50 - a_1, 50 - a_2)$. Therefore cases of $a_1 = 25 + s$ are symmetrical to $a_1 = 25 - s$ (for $s < 25$) , and I can omit all cases $a_1 > 25$ w.l.o.g.

3. If both sets A and B are empty, $a_3 = OUT$.

Akin to the sequential Hotelling game, we easily see that the SPNE is not unique, as player 2 can choose to play either $a_2 = 33$ or $a_2 = 37$ following $a_1 = 25$, as both choices are best responses. Furthermore, after many histories, the best response for player 3 is not unique off the equilibrium path. The complete SPNE, including all best responses given all histories for all players, can be found in appendix A.2.

With these action spaces and parameter choices, I chose an implementation for the lab that changes as little as possible compared to the sequential Hotelling game, while preserving the equilibrium prediction and also the intuition behind it.

The intuition for the SPNE in terms of the lab game (and therefore similarly in the sequential Hotelling game) is the following:
Consider the case of $a_1 < 25$, i.e. player 1 choosing to enter the game but not at the median location. In this case, player 2 best responds by locating to the right of the median in such a way that it is not possible for player 3 to choose a location such that $v_3 \geq max(v_1, v_2)$, and such that $v_2 > v_1$ if $a_3 = OUT$. This means that player 2 can guarantee himself a win in all subgames following $a_1 < 25$. Therefore, as player 1 can guarantee himself the higher payoff of 0.25 by choosing OUT, $a_1 < 25$ cannot be part of an SPNE.
Next, consider the case of $a_1 = 25$. Then the best response for player 2 is to play $a_2 = OUT$: If player 2 chooses any location to the left of the median or the median itself, player 3 chooses a location close to and to the right of the median and wins. Therefore, player 2 will play $a_2 = OUT$. Now player 3 can only tie with player 1 for first place by choosing $a_3 = 25$ (i.e. the median), which is profit maximizing for him. Player 1 is therefore better off choosing $a_1 = 25$ than $a_1 = OUT$ and splits the win with player 3.

So when we put this together, play along the equilibrium path consists of $a_1 = 25$, $a_2 = OUT$ and $a_3 = 25$, which is in accordance with the Osborne-Katz-conjecture, as the first and the last player enter at the median and the middle player stays out.

2.3 The Experiment

I conducted eleven sessions at the Vienna Center for Experimental Economics (VCEE) with 132 subjects. Sessions lasted about 2 hours on average. The range of earnings was between €6 and €53, with an average payment of about €26. The experiment was programmed and conducted with the software z-Tree (Fischbacher (2007)), and ORSEE (Greiner (2004)) was used for recruiting subjects.

A Parameters

The game is played over 24, 48 or 72 rounds, depending on the treatment. In each round, subjects were randomly rematched with two other subjects from the matching pool of 12 subjects to form groups of three. Each subject was then randomly assigned a position (i.e. the order in which they would act) within these groups of three subjects, with the constraint that after all rounds, every subject had been in all positions the same number of times. Subjects were assigned new positions after each round instead of keeping their positions fixed for two reasons: First, it would be unfair to subjects in middle positions in terms of payoffs, as there is a first-mover and last-mover advantage in this game; second, I thought that subjects would learn faster if they experienced the game from all player positions.

Subjects determined the locations they wanted to choose by means of a slider that only lets them choose locations that are in their action space, and there is a button labeled "No Location" for choosing $a_i = OUT$. Subjects were able to see all actions of their two group members' previous choices in this round, i.e. player 2 sees player 1's choice when he makes his decision and player 3 sees both choices from players 1 and 2. This was clearly represented on the slider and in written form on the decision screen.

After each round, subjects saw a detailed feedback screen, indicating all chosen locations by all three group members and their respective points in this round, as well as feedback on all players' payoffs in their group in this round. A subject's total payoff in the experiment was the sum of all payoffs from all rounds in Euro, i.e. the exchange rate of points in the game to Euro was 1:1. Note that with these parameter choices losses are not possible in the experiment, as the minimum payoff a subject can get in each round is 0.05€.

B Instructions and Questionnaire

The experiment started with on-screen instructions where neutral framing was used. Instructions were followed by control questions; see Section A.3 in the appendix. After the 24, 48 or 72 game rounds, a short questionnaire concluded the experiment.[17] In addition, there were questions of the form "When you were in position 1, what did you do and why?" for all positions, as well as more subtle questions; see Section A in the appendix.

C Treatments

The baseline treatment 24R corresponds to the lab game described in Section 2.2 played over 24 rounds, the second treatment changes the tiebreaking rule and the third and fourth treatment increase the number of rounds played; this is represented in Table 1. See the next section for details.

Table 1: *Treatments*

Treatment	Rounds	Observations	Subjects	Tiebreak rule
24R	24	288	36	standard
24R+A	24	288	36	alternative
48R	48	576	36	standard
72R	72	864	24	standard

Notes: The standard tiebreak rule corresponds to the one in Section 2.2, namely that if two or more players have the same maximal number of votes, they split the prize evenly. The alternative rule is that these players enter a lottery, where one of them gets the whole prize.

[17] I elicited standard socioeconomics like age, gender, income and highest education, as well as a standard Cognitive Reflection Test (Frederick, 2005). On top of that I also used a short incentivized measure for social preferences, namely a variant of a test proposed in Thibaut and Kelley (1959).

3 Results

For simplicity I will denote choice triples by (a_1, a_2, a_3),[18] and conditional choices will be written as $a_2(a_1 = a) = c$ for player 2 and $a_3(a_1 = a, a_2 = b) = c$ for player 3.[19]

As we have already seen in footnote 16 in Section A, due to the symmetric nature of the game around the median location of 25, many choice triples are symmetric and lead to the same payoff and vote shares, and are therefore handled as the same observation. So to merge all symmetric observations together, I transform all cases of $(a_1 > 25, a_2, a_3)$ into $(50-a_1, 50-a_2, 50-a_3)$; I transform all cases of $(a_1, a_2 > 25, a_3)$ into $(50-a_1, 50-a_2, 50-a_3)$ if $a_1 \in \{OUT, 25\}$; and I transform all cases of $(a_1, a_2, a_3 > 25)$ into $(50-a_1, 50-a_2, 50-a_3)$ if $a_1 \in \{OUT, 25\}$ and $a_2 \in \{OUT, 25\}$. Basically, any observation where a player is the first to choose a location a_i that is not the median is equivalent to that player choosing location $50 - a_i$, and choices made by following players are adjusted accordingly.[20]

3.1 Treatment 24R - Baseline

In this treatment, the game described in Section 2.2 was played for 24 rounds. Three sessions with 12 subjects each were conducted, resulting in 288 observations (24 rounds x 4 groups x 3 sessions).

Figure 1 shows the distribution of locational choice triples, split into rounds 1-12 and 13-24. Note first that there is an absence of the unique SPNE outcome $(25, OUT, 25)$ in rounds 1-12, and it is observed only two times in rounds 13-24 across all three sessions. In the first 12 rounds, the observation $(17, 33, OUT)$ is most common. In this case, players 1 and 2 "corner the market" and win, which means they locate in such a way that player 3 has no possibility to

[18] For example, $(17, 33, OUT)$ means that the player in position 1 chose location 17, the player in position 2 chose location 33 after observing player 1's choice of location 17, and the player in position 3 did not choose a location.

[19] For example, if player 2 chooses location 33 after observing player 1's choice of location 17, I would write $a_2(a_1 = 17) = 33$. If player 3 then observes both these choices and chooses OUT, I would write $a_3(a_1 = 17, a_2 = 33) = OUT$.

[20] For example, $(25, 25, 23) \widehat{=} (25, 25, 27)$, $(OUT, 9, 33) \widehat{=} (OUT, 41, 17)$ and
$(25, OUT, 1) \widehat{=} (25, OUT, 49)$), where $\widehat{=}$ means that they are handled as the same observation.

enter and win the game. We can also see in Figure 1 that in rounds 13-24, $(17, 33, OUT)$ is still the most frequent observation, but we also see a significant rise of $(17, 29, OUT)$.[21] The response by player 2 of $a_2 = 29$ given $a_1 = 17$ is in line with the SPNE, as player 3 can still not find a location to win or share the win, but player 2 gets more points than player 1 and wins alone.

Overall, players in position 3 chose according to the SPNE a majority of the time (67% across all 24 rounds); this high frequency is not surprising, as player 3 has no uncertainty about the behavior of the other players. Players in position 2 chose according to the SPNE only 18% of the time. Player 1 chose according to the SPNE (i.e. $a_1 = 25$) in 28% of observations. The choice of player 1 to favor $a_1 = 17$, however, was payoff-maximizing in most cases given the behavior of players 2 and 3: The payoff of players in position 1 who choose $a_1 = 17$ is about three times higher than the payoff of those who play $a_1 = 25$; more on this in Section 3.3.

3.2 Treatment 24R+A - different tiebreaking

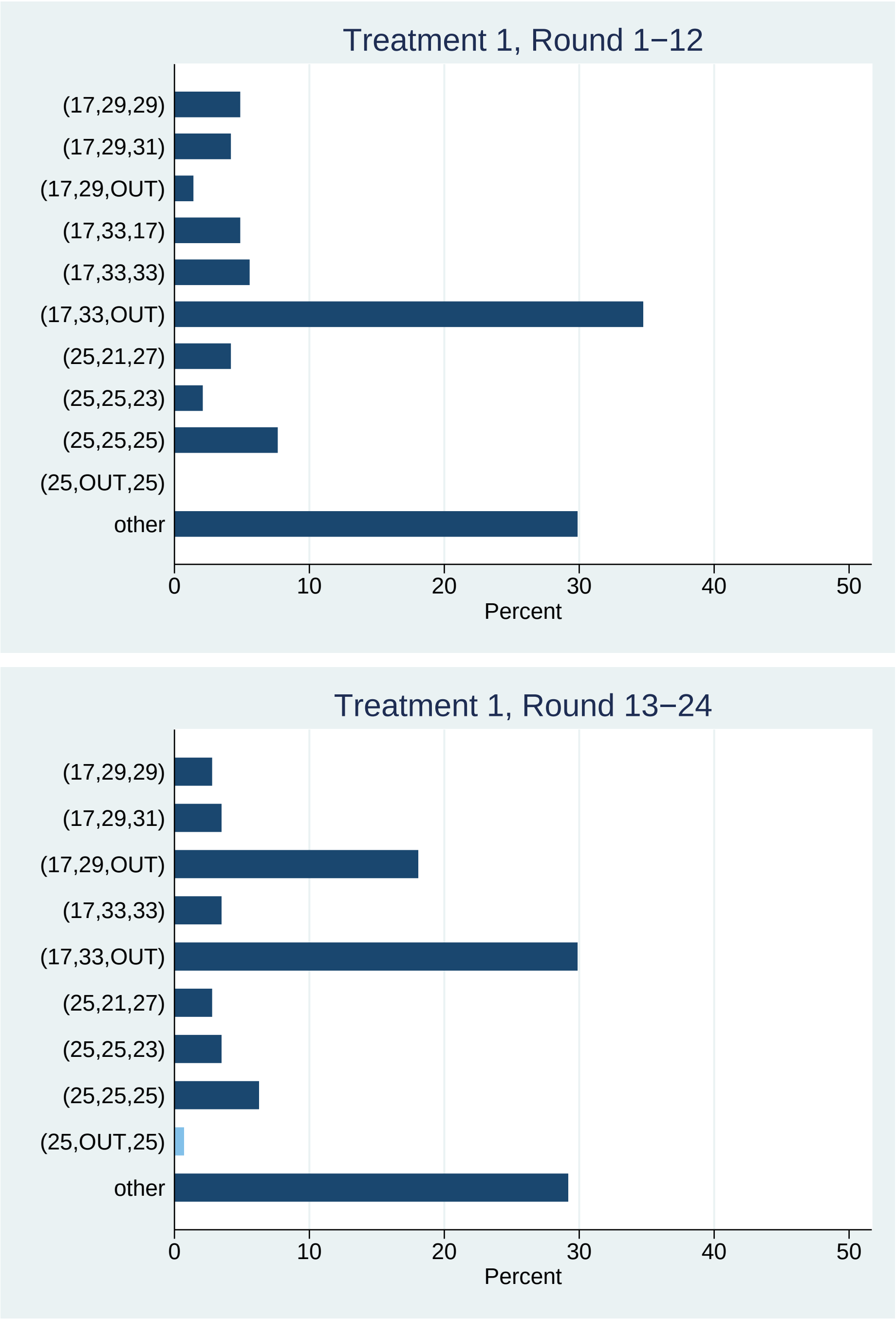
We have seen in treatment 24R that there is next to no play of the unique SPNE outcome, while $(17, 33, OUT)$ is the dominating outcome. Reciprocity might be the cause of this behavior, as it could drive player 2 to play $a_2 = 33$ given $a_1 = 17$ to share the win with player 1: If player 2 realizes that he cannot win if $a_1 = 25$ (see Section A), he might be thankful to player 1 and share the win with him while still keeping player 3 out of the game.[22]

In treatment 24R+A, the next step was to include a different tiebreaking rule than in treatment 24R to deter reciprocal behavior and give the SPNE a better shot. Consequently, the tiebreaking rule in the baseline (i.e. if two or more players have the same number of points in a round, they split the prize evenly) was changed to a rule that undermines reciprocal and egalitarian incentives: If two or more players have the same number of points at the end of a

[21] The difference in $(17, 29, OUT)$ in the first and second half of the game is significant at $p = 0.003$; OLS regression on the individual level with standard errors clustered by session, where the dependent variable is the fraction of plays of $a_2 = 29$ given $a_1 = 17$ by an individual subject, and the independent variable is a dummy for round 13-24; for the detailed specification, see footnote 26.

[22] An incentivized measure for social preferences was elicited in the experiment. The test of whether player 2's response of $a_2 = 33$ given $a_1 = 17$ occurs more frequently for subjects with high social preferences was, however, inconclusive.

Figure 1 *Learning in Treatment 24R*



Notes: Locational choice triples on the y-axis, frequency in percent on the x-axis; only 2 observations of the unique SPNE outcome; other triples with less than 20 obervations are merged in "other"; N= 288.

round, one of them gets the full prize and the other(s) get(s) nothing, with equal chances.[23] With this new tiebreaking rule, players in position 2 who choose $a_2 = 33$ given $a_1 = 17$ will now not split the prize with player 1 but rather enter a lottery for the prize. According to the standard tiebreaking rule, player 2 could (almost) be sure that he would split the prize with player 1; now, one of the two will get the whole prize, and players motivated by reciprocity will be deterred from choosing $a_2 = 33$ given $a_1 = 17$ due to the fact that if risk is involved compared to sure decisions, generous giving is significantly reduced. Brock et al. (2013) observe this behavior in dictator games, and player 2 faces a choice akin to the dictator in these games, as he can either (in a majority of cases) take the whole prize for himself, or split it with player 1.[24] On the other hand, if reciprocity does not explain this behavior, the new tiebreaking rule would not change behavior. As in the baseline, three sessions were conducted with new subjects, giving 288 observations.

Figure 2 shows that there are more observations of the SPNE outcome (highlighted) than in treatment 24R, although not significantly so (see Table 6 in the appendix). Overall choices did not change substantially compared to the baseline: The dominating presence of $(17, 33, OUT)$ and the rise in $(17, 29, OUT)$ can still be observed.[25] If anything, reciprocal behavior got stronger, as we observe fewer plays of $(17, 29, OUT)$ than in the baseline and more plays of $(25, 25, 25)$, which corresponds to a three-way tie. In fact, I detect no significant differences between treatment 24R and 24R+A in any variables. The corresponding tests based on regressions can be found in appendix A.4, Table 6.

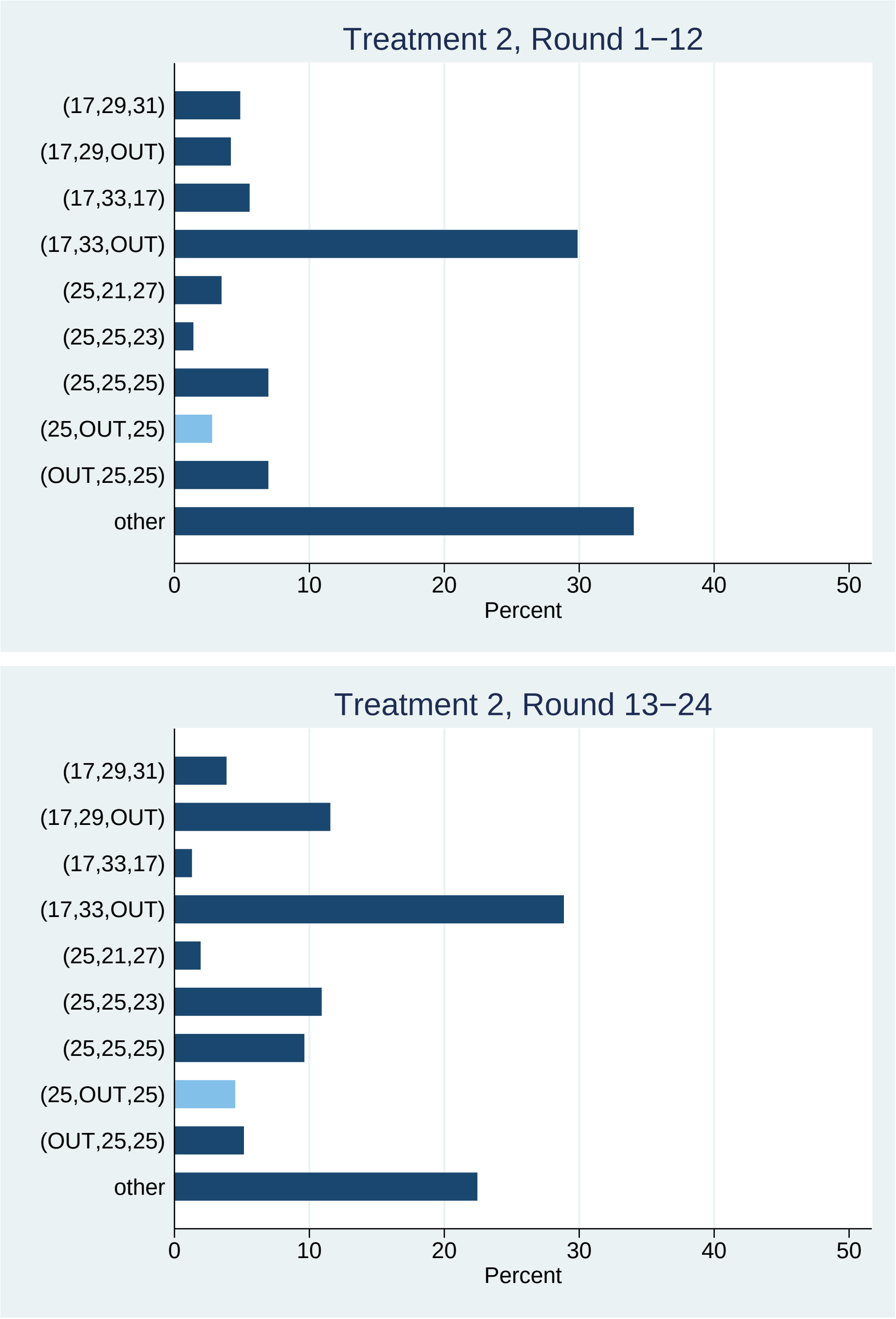
The number of plays according to the SPNE is also similar to the baseline: 28% of players in position 1, 21% of players in position 2 and 65% of players in position 3 chose actions in accordance with the SPNE. I conclude that the results from treatment 24R+A do not significantly deviate from those of treatment 24R. This, and the fact that there are no references in the questionnaires that players in position 2 play $a_2(a_1 = 17) = 33$ due to

[23] This new tiebreaking rule does not change the SPNE assuming risk neutrality, as in expectation the payoffs remain the same.

[24] Note that also in the case of $(25, 25, 25)$, which makes up about 7% of the baseline sample, reciprocal behavior might be the cause of player 3's choice to split the prize with players 1 and 2 instead of winning alone by choosing any location close to 25.

[25] As in treatment 24R, an OLS regression on the individual level clustered by session shows that the number of plays of $a_2(a_1 = 17) = 29$ rises significantly in the second half of the game, $p = 0.044$; see footnote 26 for detailed specification.

Figure 2 *Learning in Treatment 24R+A*



Notes: Locational choice triples on the y-axis, frequency in percent on the x-axis; triples with less than 20 observations are merged in "other"; unique SPNE outcome highlighted; N=288.

reciprocity, is a strong indication that prosocial preferences are unlikely to explain behavior in the treatment 24R.

3.3 Treatments 48R and 72R - Giving Subjects more time to learn

In treatment 48R the number of rounds of the game was increased to 48 and in treatment 72R to 72 rounds. As we saw no significantly different behavior with the tiebreaking rule in treatment 24R+A, the tiebreaking rule of the baseline (if two or more subjects have the same number of points at the end of a round, all of them split the prize evenly) was reinstated for treatments 48R and 72R.

The design choice to increase the length of the game follows from the observations made in the first two treatments: Learning could be observed, albeit slowly and not directly towards the SPNE. The extent of learning that can be observed in treatments 24R and 24R+A is summarized in Table 2, where I pool all data from these two treatments and compare behavior in rounds $13-24$ to behavior in rounds $1-12$. Play according to the SPNE rose from rounds $1-12$ to rounds $13-24$ for player 3 (56% vs. 66%, $p=.049$, OLS regression[26]), and players in position 2 slowly learn to best respond to $a_1=17$ with $a_2(a_1=17)=29$ (12% vs. 23%, $p=.021$, OLS regression, for specification see footnote 26). No significant learning can be observed for player 1, but this is not surprising as the subjects face a backward induction problem, so it makes sense that learning starts with players 2 and 3.

[26] The regression that is used for testing has the form $Y_{i,j}=\beta_0+\beta_1*Dummy+\epsilon_{i,j}$, where Y is the dependent variable of interest and $Dummy$ represents a dummy variable to be tested that takes values $j=\{0,1\}$, while i is the subject index. The dependent variable is a mean that is calculated individually for each subject i, and standard errors are clustered by session. The null hypothesis is $\beta_1=0$.
For example, if I want to test whether the frequency of SPNE play by player 1 (i.e. $a_1=25$) increased in rounds $13-24$ compared to rounds $1-12$, I calculate $Y_{i,0}$ and $Y_{i,1}$ for all i, where $Y_{i,j}=$(number of times subject i played $a_1=25$)/(number of times subject i was in position 1), and where $j=1$ (0) if the round number is $13-24$ $(1-12)$. I then run the above regression with the dummy variable taking value 1 if the round number is $13-24$, and 0 otherwise.
If I want to test treatment differences, I construct a treatment dummy. For dependent variables which are conditional, e.g. $a_2(a_1=17)=29$, $Y_{i,j}=$(number of times subject i played $a_2(a_1=17)=29$)/(number of times subject i was in position 2 and $a_1=17$). To check SPNE play in general for players 2 (or 3), the dependent variable is $Y_{i,j}=$(number of times subject i played a best response)/(number of times subject i was in position 2 (or 3)); best responses can be found in Table 5 in the appendix.

Table 2: *Learning in Treatments 24R and 24R+A*

	Rounds 1-12	Rounds 13-24	p-value
SPNE play by Player 1	21%	28%	.259
SPNE play by Player 2	17%	21%	.442
SPNE play by Player 3	56%	66%	.049
$a_2(a_1 = 17) = 29$	12%	23%	.021

Notes: P-value from an OLS regression, for detailed specification see footnote 26; observations from treatments 24R and 24R+A pooled; N=576.

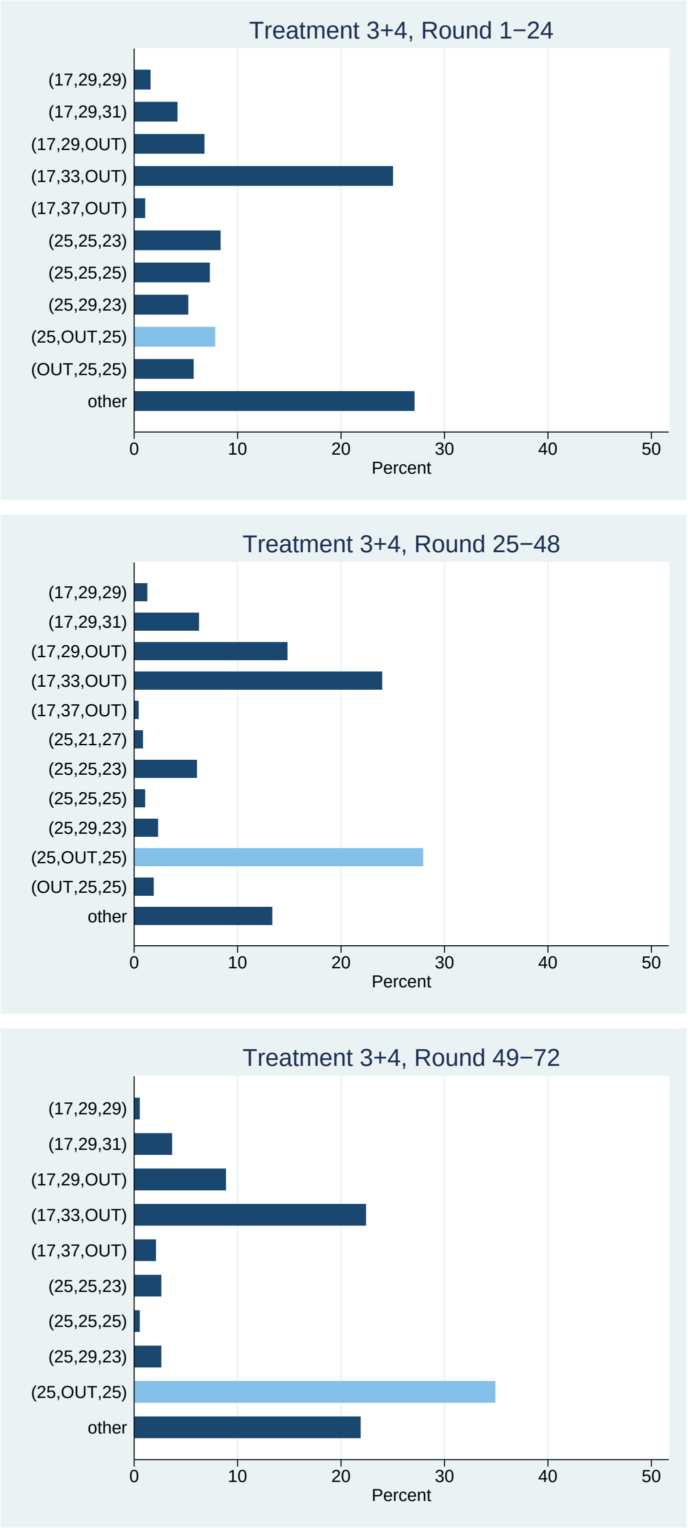
With time, more learning might be possible, so an increase in rounds appears to be the next reasonable step. New subjects were employed, and three sessions were conducted with a length of 48 rounds and 2 sessions with a length of 72 rounds.[27,28]

Figure 3 shows results for the last two treatments, where we first see that the results for the first 24 rounds are similar to those of the first two treatments. In rounds $25-48$, however, behavior changes substantially: The most frequent observation is the unique SPNE outcome with about 28%, and this number increases to 35% for rounds 49-72 in treatment 72R, while $(17, 33, OUT)$ is only the second most frequent observation.

[27] For rounds 1-24 and 25-48, 480 observations were collected from treatments 48R and 72R, while 192 observations were collected for rounds 49-72 from treatment 72R.

[28] From here on out the data from treatments 48R and 72R is pooled for rounds $1-48$ because I detect no significant differences in any variables between these treatments in rounds $1-24$ and in rounds $25-48$; see Tables 7 and 8 in appendix A.4.

Figure 3 *Learning in Treatments 48R and 72R*



Notes: Locational choice triples on the y-axis, frequency in percent on the x-axis; triples with less than 20 obervations are merged in "other"; unique SPNE outcome highlighted; N=480 for rounds 1-48 and N=192 for rounds 49-72.

Concerning differences over time, I find that in rounds $25 - 48$ the unique SPNE outcome is played significantly more often ($p = 0.049$, OLS regression, see footnote 26 for specification) than in rounds $1 - 24$. In rounds $49 - 72$, there is again a rise in $(25, OUT, 25)$ compared to rounds $25 - 48$, but the rise is only weakly significant ($p = 0.055$, OLS regression, see footnote 26 for specification); see Table 9 in appendix A.4 for both regressions. The frequency of SPNE play also rises over time, as we can see in Table 3, but the most substantial change seems to occur in rounds $25 - 48$; this will be explored further in the next section.

Table 3: *Learning over time*

SPNE play by	Rounds 1-24	Rounds 25-48	Rounds 49-72
Player 1	28%	41%	44%
Player 2	23%	43%	49%
Player 3	63%	84%	84%

Notes: All treatments pooled; N=1056 for rounds 1-24, N= 480 for rounds 25-48, N= 192 for rounds 49-72.

3.4 A closer Look at Learning

With the information gained in Sections 3.1 to 3.3, as well as utilizing insights gained from the questionnaires, there is evidence that learning to play the SPNE outcome is a multi-step process in this game, which can be observed across all sessions at variable speeds. I will explain these stages in detail in this section, and show which steps lead from the players' first play of $(17, 33, OUT)$ to the unique SPNE outcome as the most frequent observation in the end.[29]

Step 1. Player 1 and 2 corner the market: $(17, 33, OUT)$

As we have seen in all treatments in Sections 3.1, 3.2 and 3.3, in early rounds of the game $(17, 33, OUT)$ is the most frequent observation. Intuitively, this means that the players in position 1 and 2 share the win and make it impossible for player 3 to enter the game profitably, thereby "cornering the market". In the first 24 rounds, 59% of players in position 1 choose

[29] For the following analysis, I pool the data from all treatments, as OLS regressions show no significant differences across treatments; see Section A.4 in the appendix.

$a_1 = 17$, and 55% of players in position 2 respond by playing $a_2 = 33$ given $a_1 = 17$ (which is not a best response according to the SPNE), while players in position 3 make payoff-maximizing choices in a majority of cases (65%). Across all sessions, $(17, 33, OUT)$ is played 29% of the time in rounds 1-24.

We learn from the questionnaires of treatments 24R and 24R+A, where the number of rounds was 24 and therefore relatively low, that position 3 is seen as the most powerful and "easy to play" by nearly all of the subjects, as there is no uncertainty about other players' behavior when player 3 makes his decision. So it perhaps comes as no surprise that players in position 3 choose the payoff-maximizing $a_3 = OUT$ after $a_1 = 17$ and $a_2(a_1 = 17) = 33$ as their most frequent action. Additionally, choosing $a_2 = 33$ after $a_1 = 17$ in position 2 was believed to be the payoff-maximizing play by about half of the subjects.[30] Therefore, when players in position 1 learned that by playing $a_1 = 17$ player 2 would react by playing $a_2 = 33$ and player 3 would stay out of the game, they continued with this strategy.

About one third of players in position 1 do not believe entering at the median to be profitable according to the questionnaires, as they believe the other two players would respond by choosing locations close to player 1. About 20% of players in position 1 also say that they felt "lost" or "confused", as they had to predict the following players' behavior. And indeed, while $a_1 = 17$ is not according to the SPNE, in terms of payoff it is favorable: Over all treatments, players in position 1 who play $a_1 = 17$ earn 239% more than those who play $a_1 = 25$ in rounds $1 - 24$.

To sum up, in the beginning of the game a majority of players in position 1 play $a_1 = 17$, and most players in position 2 respond by playing $a_2(a_1 = 17) = 33$. Player 3 then has an easy choice to make, as there is no profitable way for him to enter the game and win, and therefore is the only player that acts according to the SPNE. Deviations from this strategy also do not pay off in the short term for player 1, as $a_1 = 17$ is far more profitable than $a_1 = 25$ in early rounds of the game. Players in position 2 have a hard time to calculate the best response to $a_1 = 17$, and given that they share a win with player 1 if they play $a_2(a_1 = 17) = 33$, a majority of players in position 2 choose to continue with this strategy. These considerations

[30] When we look at the questionnaires from treatments 24R and 24R+A, roughly 12% of subjects state that they had a hard time calculating the best response to $a_1 = 17$ when they were in position 2, and 9% of subjects indicate that they were able to figure out that the best response to $a_1 = 17$ is $a_2 = 29$, but were unsure whether to make this choice as player 2 due to uncertainty about player 3's behavior.

put together explain that $(17, 33, OUT)$ is the most frequent outcome in early rounds of the game.

Step 2. Player 2 starts to best-respond: $(17, 29, OUT)$

As we have seen in Sections 3.1 and 3.2, in rounds 13-24 the number of players in position 2 responding to $a_1 = 17$ with $a_2 = 29$ (which is indeed their best response) rises significantly, while players in position 3 still choose the payoff-maximizing option of $a_3 = OUT$ a majority of the time (65%) given $a_2(a_1 = 17) = 29$. The payoff of players in position 2 consequently rises in rounds 13-24 compared to rounds 1-12 by 20.2% $(p = 0.038)$[31], as player 2 more often wins alone rather than share the win with player 1.

Step 3. Player 1 chooses the median: $(25, a_2, a_3)$:

As players in position 1 start to lose more frequently given that player 2 best responds more often, a significant rise of $a_1 = 25$ can be observed after round 24 $(p = 0.022)$.[32] What triggers this change? A location closer to the edge, i.e. $a_1 < 17$, is almost never chosen by player 1 (which is correct, as players that do choose these locations win less than 10% of cases). If a player deviates from $a_1 = 17$, some players (10% in rounds $13 - 48$) choose $a_1 = OUT$ and stay out of the game completely, thinking that there is no possible location to win. However, most deviations from $a_1 = 17$ occur to $a_1 = 25$ (84% in rounds $13-48$), which is in accordance with the SPNE.

These deviations from $a_1 = 17$ occur even though it is still more profitable for player 1 to play $a_1 = 17$ compared to $a_1 = 25$ for most of the game (by 177% in rounds $13 - 24$, and by 48% in rounds $25 - 48$). However, in rounds 49-72, enough players in position 2 best respond, so it gets more profitable for player 1 to play $a_1 = 25$ over $a_1 = 17$ by about 8%.

Step 4. The unique SPNE outcome is the most frequent observation: $(25, OUT, 25)$

As player 1 shifts his behavior towards $a_1 = 25$, the unique SPNE outcome is not played right

[31] OLS regression on the individual level clustered by session across all treatments. Dependent variable is the mean payoff when in position 2, independent variable is a dummy for rounds 1-12 vs. 13-24; see footnote 26 for detailed specification.

[32] OLS regression on the individual level clustered by session. Dependent variable is the frequency of choosing $a_1 = 25$ when in position 1, independent variable is a dummy for rounds 25-72 vs. 1-24; see footnote 26 for detailed specification.

away. In the first 24 rounds, only 22% of players in position 2 play $a_2 = OUT$ after $a_1 = 25$, which would be according to the SPNE. This number rises significantly to 67% in rounds $25-48$ and to 85% in rounds $49-72$ ($p = .043$, OLS regression with a dummy for rounds $25-48$ vs. $1-24$, see footnote 26 for specification).

After the early stages of the game, however, as we have seen in Section 3.3, $(25, OUT, 25)$ is the most frequent observation overall in rounds $25-72$, and the rise in plays of the unique SPNE outcome is significant compared to rounds $1-24$.

3.5 Summary of Results

I will now summarize why players converge towards the unique SPNE outcome, given that initial play consists mainly of $(17, 33, OUT)$. All deviations from this strategy are towards the SPNE: First, player 2 best responds to $a_1 = 17$, and then as player 1 loses more frequently he deviates to the SPNE action of $a_1 = 25$ more often. As more players in position 1 play $a_1 = 25$, players in postion 2 also learn over time to best respond, and when they do, the unique SPNE outcome can emerge as the most frequent observation.

However, it is player 2's actions that are especially crucial for the emergence of the unique SPNE outcome. Seeing that player 3 is best responding in a majority of cases, and player 1 is profit maximizing (as $a_1 = 17$ is on average far more profitable than $a_1 = 25$ given the behavior of the other players, except in very late stages of the game), it is therefore player 2's behavior that changes most. In the beginning, player 2's reluctance to best respond to $a_1 = 17$ makes player 1's deviation from the SPNE profitable. In rounds $13-24$ player 2 starts to best respond more often. Given that player 3 mostly best responds to any (a_1, a_2), and even more so in later stages of the game, after players in position 1 start to play $a_1 = 25$ more frequently, it is again player 2's realization that he cannot win and should best respond to $a_1 = 25$ with $a_2 = OUT$ that drives the emergence of the unique SPNE outcome.

Table 4 shows that across all player positions and sessions, there was considerable learning: For all given variables, actions according to the SPNE increased over time, and choices that are not in accordance with the SPNE decreased. I therefore conjecture that with more time and more learning, the prevalence of the SPNE outcome would be even stronger.

Table 4: *Summary Statistics*

	Round			
frequency of	1-24	25-48	49-72	p-value
SPNE play in position 3	.65	.84	.84	.001
SPNE play in position 2	.25	.43	.49	.001
SPNE play in position 1	.30	.41	.47	.022
$a_1 = 17$	.59	.54	.44	.761
$a_2 = 33$ given $a_1 = 17$	.55	.51	.50	.535
$a_2 = 29$ given $a_1 = 17$	.31	.43	.44	.240
$a_2 = OUT$ given $a_1 = 25$	.22	.67	.85	.043
$(25, OUT, 25)$	.04	.28	.35	.007

Notes: Observations from all treatments pooled; p-value for the round-dummy of an OLS regression on the individual level clustered by session, the dependent variable is the frequency of the variable on the left, independent variable is a dummy for rounds 25-72 vs. 1-24; see footnote 26 for detailed specification.

4 Conclusion

In this paper, I report on a theoretical and experimental investigation of a 3-player sequential-entry variant of Hotelling's locational choice model (1929) that was proposed by Osborne and Kats. Despite clear predictions due to the uniqueness of the SPNE outcome, the experiment reveals that initial play in the experiment is not in accordance with the SPNE. However, after many repetitions play does converge toward the unique SPNE outcome.

As was stated in the introduction, this model can also be used to describe plurality-rule elections. It is interesting to see that even when behavior is not according to the SPNE, initial play suggests that a two-party system would emerge, so Duverger's law is robust to violations of the SPNE in this variant of Hotelling's model.

On a final note, we find that in many finitely repeated games with a unique equilibrium prediction, these predictions are systematically violated when tested in the lab or empirically, even in simpler environments than the one considered in this paper. Examples include ultimatum games (e.g. Roth et al. (1991) or Slonim and Roth (1998)), public goods games (see Ledyard (1995) and Chaudhuri (2011)) or the centipede game (e.g. McKelvey and Palfrey (1992)). Therefore, it is perhaps surprising that play converges toward the unique SPNE outcome in the sequential Hotelling game at all. In fact, with a shorter time horizon, I would have concluded that also in this complicated setting, behavior does not converge to the SPNE outcome at all, as we can never know when we are dealing with the long run, until it is here.

Acknowledgements

I thank the Vienna Center for Experimental Economics (VCEE), University of Vienna, for allowing me to run my experiment in their laboratory. Furthermore, I would like to show my gratitude to Maarten Janssen for introducing me to this problem in the first place, and to Jean-Robert Tyran and Karl Schlag for improving the paper with their comments and suggestions. Finally, I am immensely thankful to Wieland Müller for sharing his pearls of wisdom with me and for his patience over the course of this project.

References

Bandyopadhyay, S., Bhalla, M., Chatterjee, K. & Roy, J. (2017). Strategic dissent in the HotellingDowns model with sequential entry and private information. *Research in Economics*, 71(1), 51-66.

Barreda-Tarrazona, I., García-Gallego, A., Georgantzis, N., Andaluz-Funcia, J. & Gil-Sanz, A. (2011). An experiment on spatial competition with endogenous pricing. *International Journal of Industrial Organization*, 29(1), 74-83.

Brown-Kruse, J., Cronshaw, M. B. & Schenk, D. J. (1993). Theory and experiments on spatial competition. *Economic Inquiry* 31, 139-165.

Brown-Kruse, J. & Schenk, D. J. (2000). Location, cooperation and communication: An experimental examination. *International Journal of Industrial Organization* 18, 59-80.

Brock, J. M., Lange, A. & Ozbay, E. Y. (2013). Dictating the risk: Experimental evidence on giving in risky environments. *The American Economic Review*, 103(1), 415-437.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, 14(1), 47-83.

Collins, R. & Sherstyuk, K. (2000). Spatial competition with three firms: an experimental study. *Economic Inquiry*, 38(1), 73-94.

Duverger, Maurice (1954). *Political Parties*. London: Methuen.

Eiselt, H. A. & Laporte, G. (1997). Sequential location problems. *European Journal of Operational Research*, 96(2), 217-231.

Fischbacher, U. (2007): z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics*, 10(2), 171-178.

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

Greiner, B. (2004). An Online Recruitment System for Economic Experiments. In: K. Kremer, and V. Macho (Eds.), *Forschung und wissenschaftliches Rechnen 2003*, (pp. 79-93). GWDG Bericht 63, Göttingen: Ges. für Wiss. Datenverarbeitung.

Hotelling, H. (1929). Stability in Competition. *Economic Journal* 39, 41-57.

Huck, S., Müller, W. & Vriend, N. J. (2002). The east end, the west end, and king's cross: On clustering in the four-player Hotelling-game. *Economic Inquiry*, 40(2), 231-240.

Kephart, C. & Friedman, D. (2015). Continuous Differentiation: Hotelling Revisits the Lab. *Journal of the Economic Science Association* 1.2 (2015): 132-145.

Kress, D. & Pesch, E. (2012). Sequential competitive location on networks. *European Journal of Operational Research*, 217(3), 483-499.

Ledyard, O. (1995). Public goods: some experimental results. *Handbook of experimental economics*. Princeton: Princeton University Press (Chap. 2).

McKelvey, R. D. & Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica: Journal of the Econometric Society*, 803-836.

Neven, D. J. (1987). Endogenous sequential entry in a spatial model. *International Journal of Industrial Organization*, 5(4), 419-434.

Osborne, Martin J. *An introduction to game theory*. Vol. 3. No. 3. New York: Oxford University Press, 2004.

Osborne, Martin J. (1995). Spatial Models of Political Competition under Plurality Rule: A Survey of Some Explanations of the Number of Candidates and the Positions They Take. *The Canadian Journal of Economics* 28(2), 261-301.

Prescott, E. C. & Visscher, M. (1977). Sequential location among firms with foresight. *The Bell Journal of Economics*, 378-393.

Rabas, A. (2011). *Electoral competition with an arbitrary number of participants* (Master Thesis, University of Vienna).

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M. & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review*, 1068-1095.

Slonim, R. & Roth, A. E. (1998). Learning in high stakes ultimatum games: An experiment in the Slovak Republic. *Econometrica*, 569-596.

Thibaut, J. & Kelley, H. (1959). *The social psychology of groups*. New York: Wiley.

A Appendix

A.1 Proof for the SPNE of the sequential Hotelling game

Lemma 1. In any SPNE outcome in the sequential Hotelling game, no players are losing.

Proof. As players can always guarantee themselves a payoff of $\pi_i = 0$ by choosing $a_i = OUT$, a player will always have an incentive to deviate if he loses. □

Lemma 2. In any SPNE, if exactly two players enter the game, those players will enter at the median of 0.5.

Proof. Assuming that exactly two players i and j enter the game. If player i enters the game at a location $a_i < 0.5$ (w.l.o.g.), player j wins if he locates at the median itself, as j gets half the votes from $[0.5, 1]$, plus $\frac{x_j - x_i}{2}$, i.e. $v_j = 0.5 + \frac{x_j - x_i}{2}$, which is more than half of the votes. If both players locate at the median, they get the same number of votes, and if one player would deviate that player would get less votes and lose, which can never be part of an SPNE by Lemma 1. □

Lemma 3. If player 1 enters the game at $a_1 < 0.5$ and player 2 plays according to the following strategy profile $\hat{a_2}$, and player 3 plays $a_3 = OUT$, then player 2 will win.

$$\hat{a_2} \in \begin{cases} [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1] & \text{if } a_1 < \frac{1}{6} \\ [\frac{2-a_1}{3}, 1 - a_1) & \text{if } a_1 \geq \frac{1}{6} \end{cases}$$

Proof. With the above strategy $\hat{a_2}$, player 2 always gets more votes than player 1 because he locates closer to the median, i.e. $0.5 - a_1 > a_2 - 0.5$. □

Theorem 1. All subgame-perfect Nash equilibria of the sequential Hotelling game are given by:

$$a_1^* = 0.5, a_2^*(a_1) = \begin{cases} 0.5 & \text{if } a_1 = OUT \\ [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1] & \text{if } a_1 < \frac{1}{6} \\ [\frac{2-a_1}{3}, 1 - a_1) & \text{if } a_1 \geq \frac{1}{6} \text{ and } a_1 < 0.5 \\ OUT & \text{if } a_1 = 0.5, \end{cases}$$

while player 3 chooses according to the following rules:

1. If the set $A = \{a_3 | v_3 > max(v_1, v_2)\}$ is nonempty, i.e. if player 3 can attain $v_3 > max(v_1, v_2)$ by choosing some $a_3 \in [0, 1]$, he chooses one of these payoff-maximizing choices.

2. If set A is empty and the set $B = \{a_3 | v_3 = max(v_1, v_2)\}$ is nonempty, i.e. player 3 can attain $v_3 = max(v_1, v_2)$ by choosing some $a_3 \in [0, 1]$, he chooses one of them.

3. If both sets A and B are empty, $a_3 = OUT$.

Proof. As this is a backward induction problem, we start by looking at player 3's strategy. As player 3 is the last player to act, he simply goes through all possible location choices and chooses one of the payoff-maximizing choices given (a_1, a_2). Therefore, if player 3 chooses an a_3 according to the above strategy profile, a_3 is a best response to player 1 and 2's actions.[33]

As far as player 2's best responses are concerned, we have four cases, depending on the action of player 1:

Case 1: $a_1 = OUT$

Case 2: $a_1 < \frac{1}{6}$

Case 3: $a_1 \geq \frac{1}{6}$ and $a_1 < 0.5$

Case 4: $a_1 = 0.5$.

We will go through the different cases one by one.[34] Player 2 always has two goals, which he tries to fulfill in order. **Goal 1**: Anticipating player 3's best responses, player 2 first checks whether he can locate in such a way that it is not possible for player 3 to win the game, i.e. deterring player 3 from entering, while achieving $v_2 > v_1$. If goal 1 can be achieved, player 2 wins alone, which is preferable to all other outcomes. If goal 1 cannot be achieved, player 2 tries to achieve **Goal 2**: Player 2 tries to find a location such that he shares a win with the smallest number of players. Finally, if neither goal 1 nor goal 2 can be achieved by player 2, i.e. if player 2 cannot win, $a_2^* = OUT$.

Case 1: $a_1 = OUT$

In this case I show that player 2 cannot achieve goal 1, and he achieves goal 2 by playing $a_2^*(a_1 = OUT) = 0.5$.

If player 2 plays $a_2^*(a_1 = OUT) = 0.5$, player 3's best response is to also locate at the median of 0.5 and share the win with player 2 by Lemma 2. As $a_2(a_1 = OUT) \neq 0.5$ cannot be part of an SPNE by Lemma 2 and Lemma 1, and as player 2 can guarantee himself a shared win with player 3 by playing $a_2^*(a_1 = OUT) = 0.5$, $a_2^*(a_1 = OUT) = 0.5$ is the best response.

[33] Note as there are a great number of variations off the equilibrium path, player 3's strategy given all possible histories (a_1, a_2) is too big to explicitly write down here.

[34] Note that due to the symmetric nature of the game, we will omit all cases of $a_1 > 0.5$ w.l.o.g.

Case 2: $a_1 < \frac{1}{6}$

In this case I show that player 2 can achieve goal 1 by playing any $a_2^*(a_1 < \frac{1}{6}) \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1]$. This means that player 2 wins alone by deterring player 3 from entering (i.e. $a_3 = OUT$) while achieving $v_2 > v_1$ if $a_2^*(a_1 < \frac{1}{6}) \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1]$.

The analysis of Case 2 is structured as follows: First I show that if player 2 locates out of the given best response range, player 3 would win alone; therefore, any action $a_2(a_1 < \frac{1}{6})$ outside the best response range cannot be part of an SPNE by Lemma 1. Then I show that given $a_1 < \frac{1}{6}$ all actions in $[\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1]$ lead to player 2 winning alone, thereby achieving goal 1.

First note that we can exclude any $a_2(a_1 < \frac{1}{6}) < 0.5$ as best responses by player 2, as player 3 can then locate at the median and win alone.

Next, I show that any $a_2(a_1 < \frac{1}{6}) > \frac{2}{3} + a_1$ cannot be a best response, as player 3 can then find a location a_3 with $a_1 < a_3 < a_2$ and win alone. If $a_2 = \frac{2}{3} + a_1 + \epsilon$ where $\epsilon > 0$ and such that $a_2 \in (\frac{2}{3} + a_1, 1]$, the vote shares in this case are given by $v_1 = \frac{a_1+a_3}{2}$, $v_2 = 1 - \frac{a_2+a_3}{2}$ and $v_3 = \frac{a_2-a_1}{2}$. For player 3 to enter, two inequalities have to be fulfilled: $v_3 > v_1$, which holds iff $a_3 < \frac{2}{3} - a_1 + \epsilon$, and $v_3 > v_2$, which holds iff $a_3 > \frac{2}{3} - a_1 - 2\epsilon$. As the first two terms $\frac{2}{3} - a_1$ on the right hand side are the same in both inequalities, we see that as long as ϵ is positive, player 3 can enter the game and win alone. Therefore, any $a_2(a_1 < \frac{1}{6}) > \frac{2}{3} + a_1$ cannot be a best response by Lemma 1.

Now I show that $a_2(a_1 < \frac{1}{6}) < \frac{2}{3} - \frac{a_1}{3}$ cannot be a best response. Suppose $a_2 = \frac{2-a_1}{3} - \epsilon$ where ϵ is positive and such that $a_2 \in (a_1, \frac{2-a_1}{3})$. I will show that player 3 can then locate at $a_3 = \frac{2}{3} - \frac{a_1}{3}$ and win alone. The vote shares in this case are given by $v_1 = \frac{a_1+a_2}{2}$, $v_2 = \frac{a_3-a_1}{2}$ and $v_3 = 1 - \frac{a_2+a_3}{2}$. Again, two inequalities have to be fulfilled for player 3 to enter and win: $v_3 > v_1$ and $v_3 > v_2$. $v_3 > v_1$ holds iff $a_3 < \frac{2-a_1}{3} + 2\epsilon$, which simplifies to $0 < 2\epsilon$ by plugging in $a_3 = \frac{2}{3} - \frac{a_1}{3}$. $v_3 > v_2$ holds iff $a_3 < 1 + \frac{a_1-a_2}{2}$, which simplifies to $a_1 > -\frac{\epsilon}{2}$ by plugging in a_2 and a_3; both inequalities are always fulfilled. As I have shown that if $a_2(a_1 < \frac{1}{6}) = \frac{2}{3} - \frac{a_1}{3} - \epsilon$ player 3 can enter at $a_3 = \frac{2}{3} - \frac{a_1}{3}$ and win alone, any $a_2(a_1 < \frac{1}{6}) < \frac{2}{3} - \frac{a_1}{3}$ cannot be a best response by Lemma 1.

Now we have established that for all location choices by player 2 outside of $[\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1]$ given $a_1 < \frac{1}{6}$, player 3 can find a location to win alone, so these actions cannot be best responses for player 2 by Lemma 1. I proceed to show that for all location choices $a_2^*(a_1 < \frac{1}{6}) \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1]$, player 3 will play $a_3 = OUT$. It then follows from Lemma 3 that $a_2^*(a_1 <$

$\frac{1}{6}) \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1]$ is the best response correspondence for player 2 if $a_1 < \frac{1}{6}$. I show this by going through all possible location choices $a_3 \in [0, 1]$ for player 3 (Cases 2a-2e) and showing for each case that player 3 cannot win because either $v_3 > v_2$ or $v_3 > v_1$ cannot be fulfilled.[35] It then follows that player 3's best response given the history $(a_1 < \frac{1}{6}, a_2 \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1])$ is $a_3 = OUT$, as player 3 cannot win if he chooses any action other that $a_3 = OUT$.

(Case 2a) $a_3 < a_1$: To show that player 3 will not enter at $a_3 < a_1$, we have to show that player 3 loses if he enters at any a_3 with $a_3 < a_1$ given the history $(a_1 < \frac{1}{6}, a_2 \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1])$. In this case the vote shares are given by $v_1 = \frac{a_2 - a_3}{2}$, $v_2 = 1 - \frac{a_1 + a_2}{2}$ and $v_3 = \frac{a_1 + a_3}{2}$, and therefore $v_3 > v_2$ iff $a_3 > 2 - 2a_1 - a_2$. Given player 2's best response range $a_2^*(a_1 < \frac{1}{6}) \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1]$, $a_3 > 2 - 2a_1 - a_2$ is easiest to fulfill at the lower bound for $a_2^*(a_1 < \frac{1}{6})$. Therefore, if the inequality cannot be satisfied for the case of the lower bound of a_2^*, it cannot be satisfied for the whole best response range. If we plug in the lower bound, $v_3 > v_2$ iff $a_3 > \frac{4}{3} - \frac{5a_1}{3}$. As a_1 is bounded from above by $\frac{1}{6}$, if we were to plug in $a_1 = \frac{1}{6}$, then $v_3 > v_2$ iff $a_3 > \frac{19}{18}$, which can never be satisfied, so player 3 will not enter in case 2a.

(Case 2b) $a_3 = a_1$: To show that player 3 will not enter at $a_3 = a_1$, we have to show that player 3 loses if he enters at $a_3 = a_1$ given the history $(a_1 < \frac{1}{6}, a_2 \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1])$. In this case $v_1 = v_3 = \frac{a_1 + a_2}{4}$ and $v_2 = 1 - \frac{a_1 + a_2}{2}$, and $v_3 > v_2$ iff $a_3 > \frac{4}{3} - a_2$. Given player 2's best response range, this inequality is easiest to fulfill at the upper bound for $a_2^*(a_1 < \frac{1}{6})$. Therefore, if the inequality cannot be satisfied for the case of the upper bound of a_2^*, it cannot be satisfied for the whole best response range. If we plug in the upper bound, $v_3 > v_2$ iff $a_3 > \frac{1}{3}$, which can never be true as $a_3 = a_1 < \frac{1}{6}$, so player 3 cannot win in case 2b.

(Case 2c) $a_1 < a_3 < a_2$: To show that player 3 will not choose any a_3 with $a_1 < a_3 < a_2$, we have to show that player 3 loses if he enters at any a_3 with $a_1 < a_3 < a_2$ given the history $(a_1 < \frac{1}{6}, a_2 \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1])$. In this case, the vote shares are given by $v_1 = \frac{a_1 + a_3}{2}$, $v_2 = 1 - \frac{a_2 + a_3}{2}$ and $v_3 = \frac{a_2 - a_1}{2}$. For player 3 to enter, two inequalities have to be fulfilled. $v_3 > v_1$ holds iff $2a_1 < a_2 - a_3$ and $v_3 > v_2$ holds iff $a_1 < 2a_2 + a_3 - 2$. To see that both of these inequalities cannot be fulfilled at the same time, suppose $a_2 = \frac{2}{3} + a_1 - \epsilon$, where $\epsilon \geq 0$ and such that $a_2 \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1]$. The two inequalities then simplify to $a_3 < \frac{2}{3} - a_1 - \epsilon$ and $a_3 > \frac{2}{3} - a_1 + 2\epsilon$. We see that, as the first two terms $\frac{2}{3} - a_1$ on the right hand side are the same, as long as $\epsilon \geq 0$, both inequalities cannot be fulfilled at the same time. It follows that

[35] Note that in cases 2a-2e we have to derive player 3's best responses because we did not explicitly write down player 3's strategy given all possible histories.

player 3 cannot win in case 2c.
(Case 2d) $a_3 = a_2$: To show that player 3 will not enter at $a_3 = a_2$, we have to show that player 3 loses if he enters at $a_3 = a_2$ given the history $(a_1 < \frac{1}{6}, a_2 \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1])$. In this case the vote shares are given by $v_1 = \frac{a_1+a_2}{2}$ and $v_2 = v_3 = \frac{1}{2} - \frac{a_1+a_2}{4}$, and $v_3 > v_1$ holds iff $a_2 < \frac{2}{3} - a_1$. This inequality can never be fulfilled as $a_1 < \frac{1}{6}$ and player 2's best response range is bounded from below by $\frac{2}{3} - \frac{a_1}{3}$, so player 3 cannot win in case 2d.
(Case 2e) $a_3 > a_2$: To show that player 3 will not choose any a_3 with $a_3 > a_2$, we have to show that player 3 loses if he enters at any a_3 with $a_3 > a_2$ given the history $(a_1 < \frac{1}{6}, a_2 \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1])$. In this case the vote shares are given by $v_1 = \frac{a_1+a_2}{2}$, $v_2 = \frac{a_3-a_1}{2}$ and $v_3 = 1 - \frac{a_2+a_3}{2}$, and therefore $v_3 > v_2$ iff $a_3 < 1 + \frac{a_1-a_2}{2}$. As by assumption $a_3 > a_2$, if the right hand side of inequality $a_3 < 1 + \frac{a_1-a_2}{2}$ should equal a_2, player 3 cannot find a location such that $v_3 > v_2$ is fulfilled. So, if we solve the equation $a_2 = 1 + \frac{a_1-a_2}{2}$ for a_2, we get the lower bound for player 2's best response range, $a_2 = \frac{2}{3} - \frac{a_1}{3}$. As a_2 is substracted in the inequality $a_3 < 1 + \frac{a_1-a_2}{2}$, if $a_2 < \frac{2}{3} - \frac{a_1}{3}$, player 3 can find a location such that both $a_3 > a_2$ and $a_2 < 1 + \frac{a_1-a_2}{2}$ are fulfilled. However, if $a_2 \geq \frac{2}{3} - \frac{a_1}{3}$, player 3 cannot find a location such that both $a_3 > a_2$ and $a_2 < 1 + \frac{a_1-a_2}{2}$ are fulfilled. Therefore, player 3 will not enter the game at an a_3 with $a_3 > a_2$ after the history $(a_1 < \frac{1}{6}, a_2 \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1])$.

To sum up, player 2 loses if he plays any $a_2^*(a_1 < \frac{1}{6}) \notin [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1]$, and player 2 can deter player 3 from entering the game if he plays any $a_2^*(a_1 < \frac{1}{6}) \in [\frac{2}{3} - \frac{a_1}{3}, \frac{2}{3} + a_1]$. Player 2 therefore wins alone with this best response range by Lemma 3, fulfilling goal 1.

Case 3: $\frac{1}{6} \leq a_1 < 0.5$
In this case I show that player 2 can achieve goal 1 by playing any $a_2^*(\frac{1}{6} \leq a_1 < 0.5) \in [\frac{2-a_1}{3}, 1 - a_1)$.[36] This means that player 2 wins alone by deterring player 3 from entering (i.e. $a_3 = OUT$) while achieving $v_2 > v_1$ if $a_2^*(\frac{1}{6} \leq a_1 < 0.5) \in [\frac{2-a_1}{3}, 1 - a_1)$.
The analysis of case 3 is structured similarly to case 2: First I show that if player 2 locates out of the given best response range, player 2 will either lose or tie for the win with player 1. Then I show that given $\frac{1}{6} \leq a_1 < 0.5$ all actions in $[\frac{2-a_1}{3}, 1 - a_1)$ lead to player 2 winning alone, thereby achieving goal 1.

First note that we can exclude any $a_2(\frac{1}{6} \leq a_1 < 0.5) < 0.5$ as best responses by player 2, as player 3 can then locate at the median and win alone.

[36] Again, I omit the case of $a_1 > 0.5$ due to symmetry w.l.o.g.

Next, I show that any $a_2(\frac{1}{6} \leq a_1 < 0.5) \geq 1 - a_1$ cannot be a best response. This stems from the fact that even if player 3 does not enter at all, player 2 will split the win with player 3 (in case of $a_2(\frac{1}{6} \leq a_1 < 0.5) = 1 - a_1$) or lose (in case of $a_2(\frac{1}{6} \leq a_1 < 0.5) > 1 - a_1$), as player 1 would then be located as close or closer to the median as player 2, thereby gaining the same number of votes or more votes than player 2. This would in the best case fulfill goal 2 for player 2 (splitting the win with player 1), but as we are about to see, player 2 can fulfill goal 1 by playing any $a_2^*(\frac{1}{6} \leq a_1 < 0.5) \in [\frac{2-a_1}{3}, 1 - a_1)$, so any $a_2(\frac{1}{6} \leq a_1 < 0.5) \geq 1 - a_1$ cannot be a best response given $\frac{1}{6} \leq a_1 < 0.5$.

Now I show that $a_2(\frac{1}{6} \leq a_1 < 0.5) < \frac{2-a_1}{3}$ cannot be a best response. Suppose $\frac{1}{6} \leq a_1 < 0.5$ and $a_2 = \frac{2-a_1}{3} - \epsilon$ where ϵ is positive and such that $a_2 \in (a_1, \frac{2-a_1}{3})$. I will show that player 3 can then locate at $a_3 = \frac{2}{3} - \frac{a_1}{3}$ and win alone. The vote shares in this case are given by $v_1 = \frac{a_1+a_2}{2}$, $v_2 = \frac{a_3-a_1}{2}$ and $v_3 = 1 - \frac{a_2+a_3}{2}$. Two inequalities have to be fulfilled for player 3 to enter and win: $v_3 > v_1$ and $v_3 > v_2$. $v_3 > v_2$ holds iff $a_3 < 1 + \frac{a_1-a_2}{2}$, which simplifies to $a_1 > -\frac{\epsilon}{2}$ by plugging in a_2 and a_3, which is always satisfied. $v_3 > v_1$ holds iff $a_3 > -2 + a_1 + 2a_2$, which simplifies to $0 < 2\epsilon$ by plugging in a_2 and a_3, which is also always satisfied. As I have shown that if $a_2(\frac{1}{6} \leq a_1 < 0.5) = \frac{2-a_1}{3} - \epsilon$ player 3 can enter at $a_3 = \frac{2}{3} - \frac{a_1}{3}$ and win alone, any $a_2(\frac{1}{6} \leq a_1 < 0.5) < \frac{2-a_1}{3}$ cannot be a best response by Lemma 1.

Now we have established that for all location choices by player 2 outside of $[\frac{2-a_1}{3}, 1 - a_1)$ given $\frac{1}{6} \leq a_1 < 0.5$, player 2 will lose or split the win with player 1, so these actions cannot be best responses for player 2 as he can achieve goal 1. I proceed to show that for all location choices $a_2^*(\frac{1}{6} \leq a_1 < 0.5) \in [\frac{2-a_1}{3}, 1 - a_1)$, player 3 will play $a_3 = OUT$. It then follows from Lemma 3 that $a_2^*(\frac{1}{6} \leq a_1 < 0.5) \in [\frac{2-a_1}{3}, 1 - a_1)$ is the best response correspondence for player 2 if $\frac{1}{6} \leq a_1 < 0.5$ because player 2 wins alone and thereby fulfills goal 1. I show this by going through all possible location choices $a_3 \in [0, 1]$ for player 3 (Cases 3a-3e) and showing for each case that player 3 cannot win because either $v_3 > v_2$ or $v_3 > v_1$ cannot be fulfilled.[37] It then follows that player 3's best response given the history $(\frac{1}{6} \leq a_1 < 0.5, a_2 \in [\frac{2-a_1}{3}, 1 - a_1))$ is $a_3 = OUT$, as player 3 cannot win if he chooses any action other that $a_3 = OUT$.

(Case 3a) $a_3 < a_1$: To show that player 3 will not enter at $a_3 < a_1$, we have to show that player 3 loses if he enters at any a_3 with $a_3 < a_1$ given the history $(\frac{1}{6} \leq a_1 < 0.5, a_2 \in$

[37] Note that in cases 2a-2e we have to derive player 3's best responses because we did not explicitly write down player 3's strategy given all possible histories.

$[\frac{2-a_1}{3}, 1-a_1))$. In this case the vote shares are given by $v_1 = \frac{a_2-a_3}{2}$, $v_2 = 1 - \frac{a_1+a_2}{2}$ and $v_3 = \frac{a_1+a_3}{2}$, and therefore $v_3 > v_2$ iff $a_3 > 2 - 2a_1 - a_2$. Given player 2's best response range $a_2^*(\frac{1}{6} \leq a_1 < 0.5) \in [\frac{2-a_1}{3}, 1-a_1)$, $a_3 > 2 - 2a_1 - a_2$ is easiest to fulfill at the lower bound for $a_2^*(a_1 < \frac{1}{6})$. Therefore, if the inequality cannot be satisfied for the case of the lower bound of a_2^*, it cannot be satisfied for the whole best response range. If we plug in the lower bound, $v_3 > v_2$ iff $a_3 > \frac{4}{3} - \frac{5a_1}{3}$. As a_1 is bounded from above by $\frac{1}{2}$, if we were to plug in $a_1 = \frac{1}{2}$, then $v_3 > v_2$ iff $a_3 > \frac{1}{2}$, which of course can never be satisfied as by assumption $a_3 < a_1$, so player 3 will not enter in case 3a.

(Case 3b) $a_3 = a_1$: To show that player 3 will not enter at $a_3 = a_1$, we have to show that player 3 loses if he enters at $a_3 = a_1$ given the history $(\frac{1}{6} \leq a_1 < 0.5, a_2 \in [\frac{2-a_1}{3}, 1-a_1))$. In this case $v_1 = v_3 = \frac{a_1+a_2}{4}$ and $v_2 = 1 - \frac{a_1+a_2}{2}$, and $v_3 > v_2$ iff $a_1 > \frac{4}{3} - a_2$. Suppose $a_2 = 1 - a_1 - \epsilon$, where $\epsilon > 0$ and such that $a_2 \in [\frac{2-a_1}{3}, 1-a_1)$. $a_1 > \frac{4}{3} - a_2$ then simplifies to $0 > \frac{1}{3} + \epsilon$, which can never be true, so player 3 cannot win in case 3b.

(Case 3c) $a_1 < a_3 < a_2$: To show that player 3 will not choose any a_3 with $a_1 < a_3 < a_2$, we have to show that player 3 loses if he enters at any a_3 with $a_1 < a_3 < a_2$ given the history $(\frac{1}{6} \leq a_1 < 0.5, a_2 \in [\frac{2-a_1}{3}, 1-a_1))$. In this case, the vote shares are given by $v_1 = \frac{a_1+a_3}{2}$, $v_2 = 1 - \frac{a_2+a_3}{2}$ and $v_3 = \frac{a_2-a_1}{2}$. For player 3 to enter, two inequalities have to be fulfilled. $v_3 > v_1$ holds iff $2a_1 < a_2 - a_3$ and $v_3 > v_2$ holds iff $a_1 < 2a_2 + a_3 - 2$. To see that both of these inequalities cannot be fulfilled at the same time, suppose $a_2 = 1 - a_1 - \epsilon$, where $\epsilon > 0$ and such that $a_2 \in [\frac{2}{3} - \frac{a_1}{3}, 1-a_1)$. The two inequalities then simplify to $a_3 < 1 - 3a_1 - \epsilon$ and $a_3 > 3a_1 + 2\epsilon$. Both of these inequalities are easiest to fulfill for $a_1 = \frac{1}{6}$, so if they cannot be fulfilled for $a_1 = \frac{1}{6}$ they cannot be fulfilled for the whole range of $\frac{1}{6} \leq a_1 < 0.5$. If we plug in $a_1 = \frac{1}{6}$, we get $a_3 < \frac{1}{2} - \epsilon$ and $a_3 > \frac{1}{2} + 2\epsilon$. We see that as long as ϵ is positive, the inequalities cannot be fulfilled at the same time, so it follows that player 3 cannot win in case 3c.

(Case 3d) $a_3 = a_2$: To show that player 3 will not enter at $a_3 = a_2$, we have to show that player 3 loses if he enters at $a_3 = a_2$ given the history $(\frac{1}{6} \leq a_1 < 0.5, a_2 \in [\frac{2-a_1}{3}, 1-a_1))$. In this case the vote shares are given by $v_1 = \frac{a_1+a_2}{2}$ and $v_2 = v_3 = \frac{1}{2} - \frac{a_1+a_2}{4}$, and $v_3 > v_1$ holds iff $a_2 < \frac{2}{3} - a_1$. Given $\frac{1}{6} \leq a_1 < 0.5$, this inequality is easiest to fulfill at the lower bound of a_1, so if it cannot be fulfilled for $a_1 = \frac{1}{6}$ it cannot be fulfilled for the whole range of a_1. Plugging in $a_1 = \frac{1}{6}$, $a_2 < \frac{2}{3} - a_1$ simplifies to $a_2 < \frac{1}{2}$, which can never be fulfilled as $a_2 = a_3$ and a_3 cannot be lower than $\frac{1}{2}$ within the given response range $a_2 \in [\frac{2-a_1}{3}, 1-a_1)$. Therefore, player 3 cannot win in case 2d.

(Case 3e) $a_3 > a_2$: To show that player 3 will not choose any a_3 with $a_3 > a_2$, we have to show that player 3 loses if he enters at any a_3 with $a_3 > a_2$ given the history $(\frac{1}{6} \leq a_1 < 0.5, a_2 \in [\frac{2-a_1}{3}, 1 - a_1))$. In this case, the vote shares are given by $v_1 = \frac{a_1+a_2}{2}$, $v_2 = \frac{a_3-a_1}{2}$ and $v_3 = 1 - \frac{a_2+a_3}{2}$. For player 3 to enter, $v_3 > v_1$ must be fulfilled, which holds iff $a_3 < 2 - a_1 - 2a_2$. To see that this inequality cannot be fulfilled, suppose $a_2 = \frac{2}{3} - \frac{a_1}{3} + \epsilon$, where $\epsilon \geq 0$ and such that $a_2 \in [\frac{2}{3} - \frac{a_1}{3}, 1 - a_1)$. When we plug in $a_2 = \frac{2}{3} - \frac{a_1}{3} + \epsilon$, $a_3 < 2 - a_1 - 2a_2$ simplifies to $a_3 < \frac{2}{3} - \frac{a_1}{3} - 2\epsilon$. We see that this inequality can never be fulfilled as long as ϵ is positive because by assumption $a_3 > a_2$ and a_3 would have to be lower than the lower bound of $[\frac{2-a_1}{3}, 1 - a_1)$, so it follows that player 3 cannot win in case 3e.

To sum up, player 2 loses or ties for the win if he plays any $a_2(\frac{1}{6} \leq a_1 < 0.5) \notin [\frac{2-a_1}{3}, 1 - a_1)$, and player 2 can deter player 3 from entering the game if he plays any $a_2^*(\frac{1}{6} \leq a_1 < 0.5) \in [\frac{2-a_1}{3}, 1 - a_1)$. Player 2 therefore wins alone with this best response range by Lemma 3, fulfilling goal 1.

<u>Case 4</u>: $a_1 = 0.5$

In this case I show that player 2 can fulfill neither goal 1 nor goal 2 by choosing any $a_2 \in [0, 1]$, so $a_2^*(a_1 = 0.5) = OUT$ is the best response. First, if $a_2 < 0.5$ (which also covers the case of $a_2 > 0.5$ w.l.o.g. because of the symmetric nature of the game) given $a_1 = 0.5$, player 3 can play an a_3 such that $0.5 - a_2 > a_3 - 0.5$ (i.e. he adopts a location closer to the middle) and win alone, so any $a_2 \neq 0.5$ cannot be part of an SPNE by Lemma 1. Second, if player 2 chooses $a_2(a_1 = 0.5) = 0.5$, player 3 can choose a location close to the median and win alone. Therefore, player 2 cannot win by playing any $a_2 \in [0, 1]$ given $a_1 = 0.5$, so $a_2^*(a_1 = 0.5) = OUT$ is the best response.

Finally, we derive player 1's action in an SPNE by looking at the outcomes for player 1 in cases 1-4. In case 1, player 1 chooses $a_1 = OUT$, which is preferable to the outcomes in cases 2 and 3, where player 3 does not enter and $v_2 > v_1$ by Lemma 3, so player 1 loses. However, if $a_1 = 0.5$ (case 4), player 2's best response $a_2^*(a_1 = 0.5) = OUT$. Player 3 will play $a_3 = 0.5$ given the history $(a_1 = 0.5, a_2^*(a_1 = 0.5) = OUT)$ by Lemma 2 in an SPNE, resulting in a shared win of players 1 and 3 in case 4. As a shared win with player 3 is preferable to $a_1 = OUT$, player 1's optimal action in an SPNE is $a_1^* = 0.5$. □

Corollary 1. The unique subgame-perfect Nash equilibrium outcome for the sequential Hotelling game is given by $\{a_1 = 0.5, a_2 = OUT, a_3 = 0.5\}$.

Proof. By Theorem 1, we know that $a_1^* = 0.5$ and $a_2^*(a_1^*) = OUT$, and it follows from Lemma 2 that $a_3^*(a_1^*, a_2^*) = 0.5$. As all actions according to the SPNE are unique on the equilibrium path, the SPNE outcome is therefore also unique. □

Corollary 2. Due to the symmetry of the game, assuming $a_1 \geq 0.5$, the SPNE for the sequential Hotelling game can also be written as

$$a_1^* = 0.5, \; a_2^*(a_1) = \begin{cases} 0.5 & \text{if } a_1 = OUT \\ [\frac{1}{3} - a_1, \frac{1}{3} + \frac{a_1}{3}] & \text{if } a_1 > \frac{5}{6} \\ (a_1, \frac{1+a_1}{3}] & \text{if } a_1 \leq \frac{5}{6} \, \cap a_1 > 0.5 \\ OUT & \text{if } a_1 = 0.5 \end{cases}$$

while player 3 chooses according to the following rules:

1. If the set $A = \{a_3 | v_3 > max(v_1, v_2)\}$ is nonempty, i.e. if player 3 can attain $v_3 > max(v_1, v_2)$ by choosing some $a_3 \in [0, 1]$, he chooses one of these payoff-maximizing choices.

2. If set A is empty and the set $B = \{a_3 | v_3 = max(v_1, v_2)\}$ is nonempty, i.e. player 3 can attain $v_3 = max(v_1, v_2)$ by choosing some $a_3 \in [0, 1]$, he chooses one of them.

3. If both sets A and B are empty, $a_3 = OUT$.

Proof. In Theorem 1, we assumed w.l.o.g. that $a_1 \leq 0.5$ because of the symmetry of the game around the median of 0.5. The SPNE can also be rewritten as above while assuming $a_1 \geq 0.5$, and the proof for Theorem 1 is then equivalent if we substitute for any a_i its symmetric value $1 - a_i$.[38] □

[38] I define $1 - OUT = OUT$.

A.2 The lab game

The implementation of the sequential Hotelling game in the lab necessarily uses a discrete "voter base" (i.e. the locations 1 to 49), and not a continuous one. Therefore, the action spaces of all players are drastically smaller, and can be represented in a single table. In this section I will describe this Table 5 in detail, and show that the SPNE in the lab game results in the same unique SPNE outcome qualitatively, i.e. the first and the last player enter at the median, and the second player opts out.

In Table 5 we see all possible actions a_1 by player 1 in the first column, and in column 2 we see all possible actions by player 2 given all actions by player 1, i.e. all $a_2(a_1)$.[39] In column 3 of Table 5 we see all best response correspondences by player 3 given all possible histories of a_1 and a_2, i.e. $a_3^*(a_1, a_2)$. Finally, the last column indicates winning positions for a specific sequence of moves.[40]

In Table 5 we can therefore represent the complete action spaces by players 1 and 2, while player 3's strategies $a_3^*(a_1, a_2)$ in the table are best responses to all possible histories. Best responses to player 1's actions (while accounting for player 3's best responses) by player 2 are bold, and player 1's payoff-maximizing choice of $a_1^* = 25$ is also bold.

As an example, if $a_1 = OUT$, player 2 in principle has three options of choosing $a_2(a_1 = OUT) : a_2(a_1 = OUT) = OUT$, $a_2(a_1 = OUT) \in [1, 5, \ldots, 17, 21]$ or $a_2(a_1 = OUT) = 25$. If he chooses either $a_2(a_1 = OUT) = OUT$ or $a_2(a_1 = OUT) < 25$, player 3 plays one of his best responses a_3^* given (a_1, a_2) and wins alone (as indicated by the last column). As player 2 can split the win with player 3 by playing $a_2(a_1 = OUT) = 25$, that is his best response to $a_1 = OUT$ as it maximizes his payoff in the subgame following $a_1 = OUT$.

Player 2 chooses his best responses by looking at all possible outcomes in the subgame following a specific action a_1, anticipating player 3's best responses. From these outcomes (shown in the last column) he chooses the most favorable, i.e. he first chooses an outcome where he wins alone ("2" in the last column), then an outcome where he shares the win with

[39] Note that as in the sequential Hotelling game, we omit cases of $a_1 > 25$ due to the symmetry of the game around the median. Furthermore, we omit cases $a_2 > 25$ given $a_1 = OUT$ or $a_1 = 25$, as with these histories of player 1's action, symmetry is still preserved. See footnote 16 for a more detailed explanation.

[40] As the calculations for player 3's reaction correspondences found in Table 5 are simple and lengthy, the calculations are skipped here, but are available from the author upon request.

another player (e.g. "2 & 3"), and if no such choices are available, he chooses $a_2 = OUT$ as his best response.

Similarly for player 1, as we assume that he anticipates the best responses by players 2 and 3 following his actions in any SPNE, player 1 knows that if he chooses $a_1 = OUT$ players 2 and 3 will win, if he chooses $a_1 < 25$ player 2 will win, and if he chooses $a_1 = 25$ he will split the win with player 3, so $a_1 = 25$ is his payoff-maximizing choice and therefore the only possibility for a_1^* in an SPNE.

Table 5: *SPNE for the lab case*

a_1	a_2 given a_1	a_3^* given$\{a_1, a_2\}$	Winning Position(s)
OUT	OUT	$a_3 \in X_3 \setminus OUT$	3
	$\{1, \ldots, 21\}$	$a_3 \in [a_2 + 2, \ldots, 48 - a_2]$	3
	25	25	2 & 3
1	OUT	$a_3 \in [3, \ldots, 47]$	3
	1	$a_3 \in [3, \ldots, 49]$	3
	$\{5, \ldots, 29\}$	$a_3 \in [a_2 + 2, \ldots, 47.5 - \frac{a_2}{2}]$	3
	33	33	1 & 2 & 3
	37	$a_3 \in [27, \ldots, 35]$	3
	$\{41, \ldots, 49\}$	$a_3 \in [1 + 2(50 - a_2), \ldots, a_2 - 2]$	3
9	OUT	$a_3 \in [11, \ldots, 39]$	3
	$\{1, 5\}$	$a_3 \in [11, \ldots, 42.5 + \frac{a_2}{2}]$	3
	9	$a_3 \in [11, \ldots, 49]$	3
	$\{13, \ldots, 25\}$	$a_3 \in [a_2 + 2, \ldots, 51.5 - \frac{a_2}{2}]$	3
	29	31	3
	{**33,37**}	OUT	2
	41	25	1 & 2 & 3
	45	$a_3 \in [19, \ldots, 27]$	3
	49	$a_3 \in [11, \ldots, 31]$	3
17	OUT	$a_3 \in [19, \ldots, 31]$	3
	$\{1, \ldots, 17\}$	$a_3 \in [19, \ldots, 38.5 + \frac{a_2}{2}]$	3
	21	$a_3 \in [23, \ldots, 39]$	3
	25	$a_3 \in [27, \ldots, 31]$	3
	29	OUT	2
	33	OUT	1 & 2
	$\{37, \ldots, 45\}$	OUT	1
	49	17	1 & 2 & 3
25	**OUT**	25	1 & 3
	$\{1, \ldots, 9\}$	$a_3 \in [27, \ldots, 34.5 + \frac{a_2}{2}]$	3
	$\{13, \ldots, 21\}$	$a_3 \in [27, \ldots, 48 - a_2]$	3
	25	$a_3 \in [11, \ldots, 39] \setminus \{25\}$	3

Notes: This table provides the complete action space for players 1 and 2 in the lab game, as well as best responses given all histories for player 3; best responses bold for players 1 and 2; Section A.2 describes this table in detail; note that for all sets given here, $a_3 \in X_3$ and $a_2 \in X_2$ must still be satisfied.

Implications of Table 5

It follows from Table 5 that the SPNE in the lab game is characterized as follows, which is the same SPNE we use throughout the text:

$$a_1^* = 25, \; a_2^*(a_1) = \begin{cases} 25 & \text{if } a_1 = OUT \\ 37 & \text{if } a_1 = 1 \\ \{33, 37\} & \text{if } a_1 = 9 \\ 29 & \text{if } a_1 = 17 \\ OUT & \text{if } a_1 = 25 \end{cases}$$

while player 3 chooses according to the following rule:

1. If the set $A = \{a_3 | v_3 > max(v_1, v_2)\}$ is nonempty, i.e. if player 3 can attain $v_3 > max(v_1, v_2)$ by choosing some $a_3 \in X$, he chooses one of these payoff-maximizing choices.

2. If set A above is empty and the set $B = \{a_3 | v_3 = max(v_1, v_2)\}$ is nonempty, i.e. player 3 can attain $v_3 = max(v_1, v_2)$ by choosing some $a_3 \in X$, he chooses one of them.

3. If both sets A and B are empty, $a_3 = OUT$.

The actions that are part of the SPNE are printed in bold in Table 5. The unique subgame-perfect Nash equilibrium outcome is therefore given by $\{a_1 = 25, a_2 = OUT, a_3 = 25\}$, as all actions along the equilibrium path are unique. This result is then the same as in the sequential Hotelling game. Note that the logic behind the emergence of the unique SPNE is also very similar: As player 2 can guarantee himself a win in all subgames following $a_1 < 25$ by deterring player 3 from entering, these actions by player 1 can never be part of an SPNE. In any SPNE, the only action where player 1 is winning is $a_1 = 25$, so that is player 1's unique action on the equilibrium path.

A.3 Instructions (Baseline Treatment 24R)

Welcome to this experiment. You will be asked to make a series of choices that will affect your payoff after the experiment is over. Please pay close attention to the instructions, and do not hesitate to raise your hand in case you have any questions.

Throughout the experiment, the different payment options will be listed in Euro. In the end, you will receive the exact amount you earn in Euros.

The experiment will last for 24 rounds and will be followed by a short questionnaire. In each round, you will be in a group with two other participants, and each of you will make choices sequentially to control the biggest part of a line in order to get a payment.

Over 24 rounds, you and two other participants will form a group. All three of you will be asked to choose locations on a line. In each round, the other two members of your group will be chosen randomly, meaning that you might get completely new group members or one or two from previous rounds.

A representation of this line is given on top of the screen. The numbers on the line correspond to the locations you can choose in each round. The lowest location you can choose is 1, and the highest is 49.

A cash reward will go to the participant in a group who gets the most points in a given round. To get points, you have to control locations on the line. Each location on the line is worth a point, and the locations on the edges (1 and 49) are worth half a point, bringing the total number of points each round to 48.

The rules for how to control locations are as follows:

If you have chosen the leftmost location on the line, you control all locations to the left of you, your own chosen location and all locations that are halfway between your location and the next occupied location to your right.

If you have chosen the rightmost location, you control all locations to the right of you, your own chosen location, and all locations that are halfway between your location and the next occupied location to your left.

If your location is between two other chosen locations (meaning that your chosen location is neither the leftmost not the rightmost), you control your own chosen locations, as well as all locations that are halfway between your location and the next occupied location to your right

and all locations that are halfway between your location and the next occupied location to your left.
In essence, this means that you control a location if it is closer to your location than to any other location chosen by your other group members. If a location is equally close to two chosen locations (for example location 8 if both location 7 and 9 have been chosen), the points for controlling this location are split equally.

You also have the option to not choose a location at all. Therefore, it is possible that three, two, one or no participant has chosen a location on the line in any given round.
Also, if two or more group members have chosen the same location, all points earned are split equally between them.
At the end of the instructions, examples will be provided to further illustrate the rules.

You and the other two members of your group make your location choice sequentially, so that you make your decision in one of three positions:
If you are in position 1, you will make your choice first. If you are in position 2, you will make your choice after observing the choice of the participant in position 1 in your group. If you are in position 3, you will make your choice after observing the choice of the participants in positions 1 and 2 in your group.

The locations you can choose from are different depending on your position:
If you are in position 1, you can choose from the locations 1,9,17,25,33,41,49
If you are in position 2, you can choose from the locations 1,5,9,13,.,37,41,45,49
If you are in position 3, you can choose from the locations 1,3,5,7,,43,45,47,49
This means that the later you have to make your decision, the more locations you are able to choose from.

The payoffs are as follows: Each round, you get a fixed payment of 25 Cents. If you choose a location in that round, you have to pay costs of 20 Cents. If you do not choose a location that round, you incur no costs that round.
Then, at the end of any given round, you get a payment of 2 Euros if you have the highest number of points in your group. If two or more group members have the same highest number of points, one of them gets the payment of 2 Euros randomly with equal chance, and the others

receive nothing.
Your payoff at the end of the experiment will be the sum of all payments from all rounds.

After each round, you will see a feedback screen indicating each of your group members' chosen locations, their corresponding points from controlled locations of the line, and your payment from that round.
At the beginning of each round, you will be randomly assigned two new group members as well as randomly assigned a new position. However, over the course of the experiment, you will be in each position the same number of times, meaning that you will be eight times in each of the three positions.

A Questionnaire

Describe your behavior when you were in position 1. How was your thought process behind your decisions when you were in position 1?
Describe your behavior when you were in position 2. How was your thought process behind your decisions when you were in position 2?
Describe your behavior when you were in position 3. How was your thought process behind your decisions when you were in position 3?
Imagine that you are in position 2 during the experiment, and you observe that the group member in position 1 has chosen location 17. Which location would you choose (if any), and why?
Imagine that you are in position 2 during the experiment, and you observe that the group member in position 1 has chosen location 25. Which location would you choose (if any), and why?

A.4 Regression Tests for differences between treatments

Table 6 reports on the differences between treatment 24R and 24R+A, and we see that there are at most weakly significant differences (indicated by the stars). For a detailed explanation of the regression tests used here, see footnote 26.

Table 6: *Regressions for Treatment Effects - Treatment 24R vs. Treatment 24R+A*

mean of	Treatment 24R	Constant	R^2
$a_1 = 17$	$.094$	$.614$	$.039$
$a_1 = 25$	$-.049$	$.261$	$.017$
$a_2 = 33$ given $a_1 = 17$	$.038$	$.382$	$.006$
$a_2 = 29$ given $a_1 = 17$	$.028$	$.162$	$.008$
$a_2 = 25$ given $a_1 = 25$	$-.052$	$.167$	$.039$
$a_2 = OUT$ given $a_1 = 25$	$-.036^{*}$	$.043$	$.085$
payoff-maximizing pl. 3	$.122^{*}$	$.548$	$.066$
plays according to SPNE	$-.029^{*}$	$.022$	$.011$

*Notes: Each line represents a separate regression on the individual (subject) level with standard errors clustered by session; dependent variable is the mean of the binary variable on the left per subject; see footnote 26 for detailed specification; N=72 across all regressions (2 treatments * 12 subjects per session * 3 sessions per treatment); stars are given as follows: *: p<0.10; **: p<0.05; ***: p<0.01.*

Table 7 reports on treatment differences for rounds $1 - 24$ across all four treatments. We see that there are at most weakly significant differences (indicated by the stars). The regression we use for testing has the form $Y_{i,j,k} = \beta_0 + \beta_1 * Treatment48R + \beta_2 * Treatment72R + \epsilon_{i,j,k}$, where Y is the dependent variable of interest on the left hand side, i is the subject index, j and k take values 0 or 1 (but cannot both have the value 1) and correspond to treatments 48R and 72R. Treatment 48R and Treatment 72R are dummy variables. The dependent variable is a mean that is calculated for each subject i, and standard errors are clustered by session. The null hypotheses to be tested are $\beta_1 = 0$ and $\beta_2 = 0$.

For example $Y_{3,1,0}$ would be the mean of a left hand side variable for subject 3, who was in treatment 48R, over all 24 rounds. For further examples, see footnote 26.

Table 7: *Treatment Effects for Round 1-24 - Treatment 24R and 24R+A vs. Treatment 48R and 72R: No Significant Differences*

mean of	Treatment 48R	Treatment 72R	Constant	R^2
$a_1 = 17$	$-.155$	$-.127$	$.658$	$.097$
$a_1 = 25$	$.130$	$.127$	$.238$	$.110$
$a_2 = 33$ given $a_1 = 17$	$-.205^*$	$-.087$	$.399$	$.186$
$a_2 = 29$ given $a_1 = 17$	$.047$	$-.040$	$.175$	$.044$
$a_2 = 25$ given $a_1 = 25$	$.000$	$.014$	$.142$	$.002$
$a_2 = OUT$ given $a_1 = 25$	$.120$	$.066^*$	$.023$	$.214$
payoff-maximizing pl. 3	$.075$	$.099^*$	$.609$	$.037$
plays according to SPNE	$.116$	$.096^*$	$.349$	$.121$

*Notes: Each line represents a separate regression on the individual (subject) level with standard errors clustered by session; dependent variable is the mean of the binary variable on the left per subject; see footnote 26 for detailed specification (here we have two treatment dummies except for one, and both are tested); N=132 across all regressions (4 treatments * 12 subjects per session * 3(2) sessions per treatment); stars are given as follows: *: p<0.10; **: p<0.05; ***: p<0.01.*

Table 8 reports on treatment effects between treatments 48R and 72R, and we see that there are no significant differences. For a detailed explanation of the regression tests used here, see footnote 26.

Table 8: *Treatment Effects for Rounds 25-48 - Treatment 48R vs. Treatment 72R: No Significant Differences*

mean of	Treatment 72R	Constant	R^2
$a_1 = 17$	$.034$	$.517$	$.005$
$a_1 = 25$	$.000$	$.417$	$.000$
$a_2 = 33$ given $a_1 = 17$	$.220$	$.181$	$.177$
$a_2 = 29$ given $a_1 = 17$	$-.193$	$.312$	$.236$
$a_2 = 25$ given $a_1 = 25$	$-.005$	$.073$	$.002$
$a_2 = OUT$ given $a_1 = 25$	$.005$	$.281$	$.000$
payoff-maximizing pl. 3	$.017$	$.837$	$.002$
plays according to SPNE	$-.026$	$.576$	$..003$

*Notes: Each line represents a separate regression on the individual (subject) level with standard errors clustered by session; dependent variable is the mean of the binary variable on the left per subject; see footnote 26 for detailed specification; N=60 across all regressions (12 subjects per session * 5 sessions); stars are given as follows: *: p<0.10; **: p<0.05; ***: p<0.01.*

Table 9 tests whether there is a significant rise in plays of the unique SPNE in rounds $1 - 24$ compared to rounds $25 - 48$, and in rounds $25 - 48$ compared to rounds $49 - 72$. The first difference is significant, the second one is not. For a detailed explanation of the regression tests used here, see footnote 26.

Table 9: *Regression for Learning to play* $(25, OUT, 25)$

	Round Effect	Constant	R^2	N
Rounds 1-24 vs. 25-48	$.143^{**}$	$-.081$	$.111$	228
Rounds 25-48 vs. 49-72	$.143^{*}$	$.063$	$.036$	120

Notes: Each line represents a separate regression on the individual (subject) level with standard errors clustered by session; dependent variable is fraction of times a subject was in a play of $(25, OUT, 25)$ *in any position; round effect is a dummy variable that takes the value 0 for the earlier rounds and 1 for the later rounds in each line; see footnote 26 for detailed specification; the same coefficient of .143 in both lines is not an artifact and arose by chance; stars are given as follows: *: p<0.10; **: p<0.05; ***: p<0.01.*

Underconfident Women Earn Less - A Virtual Lab Approach

Alexander Rabas* Rupert Sausgruber† Jean-Robert Tyran‡

June 23, 2017

Abstract

Women systematically sort into different jobs than men, but isolating why they do is difficult in the field. We study job choices in a clean environment which allows us to focus on the role of beliefs about one's own productivity for job sorting. Using a quasi-representative sample of the Danish population we find that highly productive women are about twice as likely to be underconfident as men, and therefore sort into jobs that pay less. We match experimental choices with income data from official registers and find that women who are highly productive but underconfident in our experiment also earn less than their equally productive peers in the field.

Keywords: Gender, Sorting, Beliefs, Experiment

JEL Classification: C91 · D31 · J16

*University of Vienna, Doctoral School of Economics Vienna, Department of Economics alexander.rabas@gmail.com

†Department of Economics, Vienna University of Economics and Business, Vienna, Austria rupert.sausgruber@wu.ac.at

‡University of Vienna and University of Copenhagen, jean-robert.tyran@univie.ac.at

1 Introduction

Women tend to sort into different occupations than men. This tendency is a strong feature of labor markets around the world (Cortes and Pan 2017). For example, women are overrepresented in people-oriented occupations and "care work" such as teaching and nursing, and women are considerably more likely to sort into the public sector than men.[1] According to Levanon and Grusky (2016), gender sorting is so pronounced that 53 percent of the employed women in the United States would have to be shifted to a different occupation to eliminate all gender segregation. Sorting is not at all neutral with respect to incomes since wages tend to be higher in male-dominated than in female-dominated occupations. According to Lordan and Pischke (2016), the average hourly wage in occupations with a majority (more than 70 percent) of male workers is about 23 percent higher than in those with majority of female workers in the US. As a consequence, sorting into occupations and industries explains a substantial part of the gender wage gap (Blau and Kahn 2017, Card et al. 2015, Bayard et al. 2003).

There are many reasons why men a women sort into different jobs. Traditionally, the literature has emphasized gender differences in discrimination, human capital accumulation or preferences over job characteristics (e.g. Azmat and Petrongolo 2014). More recently, the literature also investigates the role of risk aversion, social preferences and attitudes towards competition (see Cortes and Pan 2017 for a survey). Experimental studies have argued that preferences for competitiveness and men's overconfidence explain why women tend to be underrepresented in settings with intense competition (e.g. Niederle and Vesterlund 2007, Almås et al. 2015, Reuben et al. 2012 and 2015).

This paper shows that confidence into one's own skills is an important determinant of gender sorting into jobs in a non-competitive setting. In particular, we show in a tightly controlled setting that high-productive women who are underconfident tend to sort into jobs that are unprofitable for them. They therefore earn less than they could have, had they sorted into a job suitable for high-productive workers. We also show that this result from our experiment carries over to incomes earned in the Danish labor market. Highly productive women who we find to be underconfident under the controlled conditions of the experiment earn less in the Danish labor market than otherwise similar women who are not underconfident.

[1] In Germany, 33% of women work in the public sector compared to only 22% of men (Dohmen and Falk 2011). In the US, women are 9% more likely to work in the public sector than men (Lewis and Frank, 2002).

Our study proceeds as follows. We recruit a large, quasi-representative sample of the Danish population to participate in a web-based experiment. Participants choose between working in a job with steep vs. flat incentives. Both jobs involve the same real effort task (counting colored squares in a grid) and are paid according to a piece-rate scheme, but the job with steep incentives is profitable for highly productive workers while the job with flat incentives is profitable for low-productive workers. Thus, mistaken job sorting is costly in terms of earnings foregone. We elicit participants' beliefs about their own productivity and observe their true output. We find that highly productive women are about twice as likely to be underconfident as men, and therefore tend to sort into a job that pays less. To investigate whether this result generalizes, we match experimental choices with income data from official tax registers in Denmark. We find that women who are highly productive but underconfident in our experiment also earn less than their equally productive peers in the field. While this finding is correlational rather than causal in nature, it is remarkable that underconfidence in the stylized experimental setting predicts incomes in the field.

Cleanly isolating the role of self-confidence in occupational sorting is a daunting task in the field and fraught with difficulties even in the laboratory. Our design has been specifically chosen to isolate the effects of overconfidence on gender sorting which means we control for various confounds that may play a role in the wild. For example, we let subjects choose between two jobs that are identical in all respects except for how well they pay for performance (steep vs. flat incentives). Hence, preferences over job characteristics cannot matter for sorting. Reputational or career concerns cannot matter for sorting in our setting because they work on the task only once. Discrimination cannot play a role for sorting in our setting because workers simply choose a job - there is no labor demand in our experiment. We implement a task that is gender-neutral which means that men and women are equally productive in the task on average. Hence, gender differences in productivity cannot matter for sorting. We measure output without error and we elicit absolute beliefs about own productivity which gives us a good measure of over- and overconfidence. We elicit a number of other correlates like IQ, personality, and risk aversion to estimate whether they may matter for sorting (they don't, except for those "on the fence, i.e. those that are almost indifferent between the two jobs).

Our paper adds to the very slim literature investigating whether self-confidence[2] into one's own skills matters for occupational choice in a clean environment. Moreover, we are the first to demonstrate that mistaken sorting in a tightly controlled experimental setting correlates with earnings in the labor market. The closest matches to our paper we are aware of are Dohmen and Falk (2011) and Larkin and Leider (2012).

Dohmen and Falk (DF, 2011) experimentally study sorting when workers (students at the University of Bonn) are faced with a choice between jobs with fixed vs. variable payment (either a piece rate, a tournament, or a revenue-sharing scheme) for work. DF focus on gender but not so much on overconfidence.[3] DF find that men are more likely to sort into the piece-rate scheme than women (74% vs. 48%). This result is not entirely straightforward to interpret because the task (multiplying numbers) in DF is not gender-neutral (when forced to work for five minutes at a given piece rate, women are less productive than men). Perhaps surprisingly, and in contrast to our study, DF find that overestimating one's productivity does not predict sorting into the piece-rate scheme. DF is similar to our paper in that they provide evidence that their findings in the lab generalize to the labor market. DF estimate how the probability of working under a variable-payment scheme in the wild is related to a proxy for productivity (education), risk attitudes (a survey measure) etc. on a different set of people (drawn from the German Socio-Economic Panel). They then compare these estimates to the estimates they obtain from the tendency to choose between fixed vs. variable experimental pay schemes in the sample of students from U Bonn. Our approach is different. We evaluate external validity by observing the incomes in the (Danish) labor market of the actual participants in our experiment. We are able to do so by matching experimental choices of our non-student participants with official register data.

Larkin and Leider (LL, 2012) focus on the role of overconfidence on sorting as we do, but they do not emphasize gender and they do not attempt to check how their results compare

[2] We use the expression self-confidence as referring to (mis-)predicting one's own (absolute) performance. Accordingly, we use the expressions under- and overconfidence for workers who under- or overestimate their own true output. Moore and Healy (2008) note that the following are, somewhat confusingly, also called overconfidence in the literature: overplacement of one's performance relative to others, and excessive precision in one's beliefs (and vice versa for underconfidence).

[3] Cadsby et al. (2007) also experimentally study job sorting by providing subjects a choice of fixed vs. piece-rate scheme, but their study is not closely related to ours since they neither focus on overconfidence nor on gender.

to a different sample as in DF or to the labor market, as we do. LL is closely related to our paper in that they let subjects choose between two piece-rate schemes (as we do). LL use a multiplication task (as DF do), and use a student sample (from Harvard University) while we use a quasi-representative sample from the Danish population. LL let subjects choose between a linear piece-rate scheme and a "convex scheme in which the piece-rate increases with output. LL show that the convex incentive scheme attracts risk seeking and overconfident subjects. This result is similar to our finding that the overconfident tend to sort more into the Steep scheme. LL study learning effects (in 9 rounds of choices), to see how persistent overconfidence is, while we use a one-shot game. LL find that the highly overconfident are 45 percentage points more likely to incorrectly choose the convex scheme. We also find mistaken sorting of unproductive workers (mainly men) into Steep, but our main result concerns mistaken sorting by underconfident women. LL find no incentive effects (as we do) since subjects who are randomly allocated to the convex scheme did not perform any better than those randomly assigned the linear scheme. LL find that overconfidence in one's own performance in the task is correlated with being male and extroverted (we also find a significant effect of extraversion for those "on the fence, i.e. those that should be nearly indifferent between Flat and Steep, and we also find that the more emotionally stable sort more into Steep).

Under- and overconfidence in one's absolute skills (aka overestimation) as discussed in the studies above has received surprisingly little attention as an explanation of gender sorting into jobs. In contrast, under- and overconfidence in one's skills relative to others (aka overplacement) has received considerable attention in the experimental literature. A large literature shows that women tend to shy away from jobs in which they need to compete with others (see Niederle and Vesterlund 2011 for a survey). This result seems very robust as it has been replicated across a broad range of cultures and shown to be present already at an early age (Gneezy et al. 2003, Croson and Gneezy 2009, Gneezy and Rustichini 2004, Gneezy et al. 2009, Cason et al. 2010, Sutter and Rützler 2010, Dreber et al. 2011, Datta Gupta et al. 2013, Buser et al. 2012, Cárdenas et al. 2012, Kamas and Preston 2012, Brandts et al. 2015).

The almost exclusive focus in the experimental literature on sorting into competition rather than sorting in non-competitive settings (as we do) also has drawbacks. First, sorting into competition typically focuses on beliefs to win a tournament, which is a highly strategic situation. Quantifying the costs of mistaken sorting is not straightforward in such an environment. Second, mistaken sorting in a strategic environment will not necessarily result in losses, as payoffs depend on what competitors a player is matched to. Players therefore have to form beliefs about their performance relative to their competitors but also need to form beliefs about the beliefs of their competitors. This complication makes it difficult to tease out the role of beliefs in sorting. A recent paper by van Veldhuizen (2017) in fact sheds doubt on the role of the psychological trait of competitiveness in sorting. Third, studying sorting into tournaments provides a plausible framework to understand sorting in highly competitive situations and to explain, say, why women are underrepresented among CEOs or professorial positions in prestigious universities. However, for the bulk of sorting choices competition does not seem to be the most salient feature, as is, for example, true of many public sector jobs.

We proceed as follows. Section 2 describes the experimental design, in particular the work task, the payment scheme and experimental procedures of our virtual laboratory approach. It also explains how we isolate the effect of beliefs on sorting. Section 3 presents results. Section 3.1 shows that the task is gender neutral in the sense that men and women are equally productive. Section 3.2 shows that beliefs indeed explain sorting but that risk aversion, IQ and personality traits do not, except for those “on the fence. We show that highly productive women tend to be underconfident. Section 3.3 discusses how sorting maps into earnings. Section 4 shows that our results generalize to the Danish labor market by matching experimental data to income data from the official Danish registers. Section 5 concludes.

2 Design

Here is a broad description of our experiment. Subjects choose between two jobs which both involve exerting real effort at the same work task but have different piece rate schemes. One job pays no flat fee but rewards work at a steep rate. The other job does pay a flat fee but rewards work at a flat rate. Hence, highly productive workers earn more in the former scheme (called Steep below), and low-productive workers earn more in the latter scheme (called Flat). Earnings exclusively depend on one’s own performance, and there is no element of competition between subjects. Before they make their choices, workers indicate their expectations about how many tasks they will complete.

The rest of this section explains procedures and parameters of our online experiment and how we identify sorting. For the instructions and a sample screen, see appendix section A.5.

2.1 Work task and payment schemes

The work task we use involves counting the number of yellow squares in a 10-by-10 grid of blue squares for 15 minutes (see appendix section A.5, Figure A7, for a sample screen). Subjects could stop working at any time.

The difficulty of the task is increasing over time. In the first few tasks there are about 8 yellow squares and this number increases slowly to about 45 yellow squares after about 80 completed tasks and remains at that level afterwards. All subjects face the same screens in the same random sequence. They are made aware of the increasing difficulty of the task in the instructions, but not exactly of how fast it increases.

Subjects only proceed to the next task if they provide the correct answer, in which case a new screen with a different pattern of yellow squares appears. If not, they need to retry and answer the same screen again. There is no penalty for answering incorrectly other than time lost.

The jobs reward performance as follows:

Flat: $\pi_i = 60 + 0.5a_i$
Steep: $\pi_i = 0 + 1.5a_i$,

where π_i indicates the earnings of subject i in Danish Crowns (DKK), and a_i is the total number of correctly answered questions. By solving the two equations for a_i, we find a cut-off point at 60 tasks, corresponding to earnings of 90 DKK (approx. 12 Euro). That is, earnings are higher in Steep for the highly productive, i.e. those who complete more than 60 tasks, but are higher in Flat for the low productive (complete less than 60 tasks). We explain this fact to subjects in the instructions. Correct sorting would thus imply that the highly productive workers sort into Steep and the low-productive workers sort into Flat. Hence, two types of sorting errors can occur when this is not the case.

We have chosen these jobs (i.e. the combination of task and payoff structure) because they provide ideal conditions to test for the relevance of (mistaken) beliefs about one's own productivity in job sorting. By virtue of experimental control, we can rule out a number of other factors that may also play a role in gender sorting in the field but cannot play a role here.

This ability to control allows us to focus on beliefs as the key factor of interest in a way that is not possible in the field, as we discuss next.

First, we have chosen the counting task because it is gender-neutral, meaning that men and women are equally good at the task. Systematically different performance across genders would mean that one type of sorting error would be systematically more common for men than for women. Suppose, hypothetically, we had chosen a task in which men are systematically more productive than women such that all men complete more than 60 tasks and all women complete less than 60. In this case, overconfident men, i.e. men who think they can produce more than they effectively do, would not make a sorting mistake. Similarly, we could not detect mistaken sorting by underconfident women in this case. But this is not case in our setting. In fact, the overall output is almost exactly 60 tasks (in fact, 60.32), and there are no significant differences across genders (see section 3.1 for details). This means that our design creates ideal conditions to observe overconfidence and underconfidence, and to observe it equally likely across genders.

Second, the incentive structure we have chosen does not provide incentives to err on one side rather than the other. In fact, the costs of either type of mistaken sorting are symmetric as both payoff functions are linear. To illustrate, suppose a worker believes he would complete 60 tasks but is uncertain about whether his output will be higher or lower than that number. If the costs were higher for sorting into Steep than Flat, such a person would rationally choose Flat. Asymmetric payoff schemes would thus induce a sorting bias by design. Our design avoids any such bias, and it does so for both genders.

Third, we have chosen a design in which workers' earnings exclusively depend on their own productivity, i.e. there are no spillovers from one worker to the other, workers do not compete, and no strategic reasoning is required to make the optimal choice. Hence, beliefs about the performance of other workers and, by implication, relative productivity beliefs, do not matter for job sorting, nor do social preferences. These factors may play a role in contexts with competition between workers which have been studied in much of the literature (see Niederle and Vesterlund, 2011, for an overview). Van Veldhuizen (2017) notes that because relative and strategic concerns matter in settings with competition, it is difficult to tease out the relative power of overconfidence, risk preferences and competitiveness in explaining why women shy away from competition, and suggests a novel way to assess the importance of these factors.

2.2 Beliefs and identification of sorting effects

Our goal is to test whether mistaken beliefs about one's own performance drive job sorting, and to investigate how job choices differ by gender. Hence, we need to elicit such beliefs and compare them to actual performance. Subjects are asked to indicate how many tasks they think they will complete within 15 minutes before they make the job choice. They do so separately for both schemes. That is, they indicate how many tasks they think they will complete in case they work under Steep and how many they think they will complete under Flat. This elicitation was not incentivized to give subjects no incentive to stop working after they have reached their stated productivity in the experiment.

Subjects then choose Steep or Flat, knowing that their choice might be overruled. In fact, 50% of subjects get randomly assigned to a job, and the other 50% get the job they chose. This means that a subject has a 75% chance to work under his or her preferred job.

The reason we overrule the choice of some subjects, i.e. force them to work under the job they did not choose, is that this procedure allows us to identify sorting mistakes. Here is the intuition why this is so (for more detailed explanations see appendix A.2). Consider a worker who chose Flat and completes 50 tasks. Did this worker sort correctly into Flat or did he make a sorting error? The answer depends on how much he would have produced had he worked under Steep incentives. Suppose he would have completed 70 tasks in Steep because he reacts to the stronger marginal incentives (1.5 DKK vs. 0.5 DKK per task) in Steep vs. Flat by working harder. In this case, the choice of Flat would have been a sorting mistake because he could have earned more by choosing Steep. Now, suppose the worker would have completed also 50 tasks in Steep. In this case, he would have made no sorting mistake because his earnings are higher in Flat in this case. Hence, we need to know the counterfactual output (i.e. how much a worker would have produced in the job that he has not chosen) to identify sorting mistakes. It is exactly to obtain this counterfactual that we force some workers who chose Flat to work under Steep, and vice versa. To obtain the counterfactual, we compare the average output of workers who made the same choices but got different incentives.

Results show that workers' output does not significantly respond to the difference in marginal returns in Steep vs. Flat, i.e. we find no incentive effects, and we also do not find a gender difference in the incentive effect. Neither average output nor distributions are significantly different in forced Steep vs. chosen Flat ($p = 0.623$, Mann-Whitney-U-Test, $p = 0.625$ Kruskal-Wallis test). The results are analogous for forced Flat vs. chosen Steep ($p = 0.846$ both tests). While this non-response to a strong change in incentives might seem surprising, such non-response is the rule rather than the exception in the experimental literature (see Cappelen et al. 2013 or Eckartz et al. 2012, as well as Augusto et al. 2015, Corgnet et al. 2015 and Eckartz 2014 for discussions). The reason is that workers in laboratory real effort tasks often only have a relatively short time to work on the task and no alternative to working. Hence, they tend to work the entire the time and at full intensity even when marginal incentives are low. In fact, we find that 93% of our subjects work all 15 minutes. Therefore, there is no room to increase output at higher marginal incentives. This is not surprising, because contrary to similar experiments, the piece rate (and therefore the marginal incentive to answer an additional question) is strictly positive in both schemes.[4]

2.3 Data

A virtual Lab Approach

Our study uses the platform iLEE (internet Laboratory for Experimental Economics) which has been developed at the Department of Economics at the University of Copenhagen.[5] The virtual lab combines the advantage of lab experiment with a the possibility to recruit large samples over the web. That is, the virtual lab approach follows the standards (e.g. no deception, payment according to choices) and procedures (e.g. with respect to instructions) as in a conventional laboratory experiment but subjects make choices remotely, over the internet. In collaboration with Statistics Denmark, a Danish governmental agency, we recruit large quasi-representative samples of the Danish population and this collaboration allows us to match experimental choices with data from official registers (see Thöni et al. 2012 for a description). The ability to recruit such samples is ideal for our purpose because it allows us to closely match the composition of the general working population in Denmark which is important for demonstrating the external validity of our findings. To date, iLEE has been used to run four waves of experiments, each consisting of several modules. The experimental data reported here comes from iLEE3, the

[4] Incentive effects are found when the choice is between jobs with positive v. zero marginal incentives (e.g. Dohmen and Falk, 2011).

[5] For more information about iLEE see: http://www.econ.ku.dk/cee/iLEE

third wave of the iLEE project.

Procedural Details

A random sample of 22,207 subjects was drawn from the general voting-age population of Denmark for the first wave of iLEE, called iLEE1. Statistics Denmark (SD) sent hard-copy letters by regular mail, inviting participants to log into iLEE's webpage using an ID number generated by SD for this purpose and to complete the study within a week. Upon login, subjects were informed that the study would take approximately 50 minutes and that they would only be paid if they completed the entire study. The study was conducted in Danish. The data was then made available to us in a fully anonymous format and was matched by DS with detailed data from official Danish registers. The first wave was rolled out in 2008 and was followed by one wave per year. For further waves of iLEE, only subjects who had participated in the first wave were invited.

For the third wave of iLEE of which our experiment was part, 2,244 invitations were sent out; all of these recipients had completed iLEE1 and more than half of these had also completed iLEE2. iLEE3 consisted of several modules and 715 participants completed the particular experiment reported here.[6] As a consequence, we have substantial attrition from iLEE1 to iLEE3 which explains why the sample of subjects participating in the experiment reported hereis only quasi-representative, i.e. is representative of the Danish population among some dimensions but not fully representative in others. For example, our sample is reasonably close to the proportions in the Danish population with respect to gender (47.1% vs. 50.2% in the population), the young (18-30 years: 16.2% vs. 18.5%, 33 to 40 years: 25.9% vs. 29.1%) and the old (60 to 80 years: 21.2% vs. 25.3%). But there is clear overrepresentation of the middle aged (45 to 59 years: 38.5% vs. 27.0%), the highly educated (long tertiary education: 18.9% vs. 7.1%), and high-income earners (more than 400,000 DKK p.a. 29.3% vs. 15.0%). See appendix section A.1, Table A1, for further details.

Our analysis uses various measures which were elicited in different waves. The personality measures are derived from the short version of the Big5 personality test in McCrae and Costa

[6] ILEE3 had 1,046 participants in total, but not all of them took part in the particular experiment discussed in this paper due to the modular structure of iLEE.

(2003).[7] We elicited risk aversion via the original Holt and Laury method (Holt and Laury 2002). Our measure of IQ is based on the I-S-T 2000R intelligence structure test, which, in turn, is based on Raven's progressive matrices (Raven, 2003); subjects have 10 minutes to solve 20 puzzles, and our IQ measure is the number of correct answers given. Neither of these tests was incentivized.

3 Results

Section 3.1 shows that there are no gender differences in output. However, there is significant mistaken gender sorting in our experiment, as highly productive women choose Flat too often, given their productivity. Section 3.2 shows that the explanation for this behavior are productivity beliefs: Women are on average less confident about their productivity than men, especially women of high ability. Section 3.3 shows that women tend to lose money because they are underconfident and men because they are overconfident. Section 4 shows that that our main result translates to field data: Underconfident women in the experiment earn less in the field than non-underconfident women.

3.1 Sorting in a gender-neutral task

Figure 1 shows that the task was gender-neutral in the sense that men and women were equally productive on average. The mean output over all participants is 60.32 tasks which is very close to the indifference point of 60 tasks. Having observations centered on this number is ideal for our purposes because it makes both types of sorting errors equally likely. However, having a large mass of observations close to the indifference point also means that errors are likely to be common (They indeed are, as we show in the next paragraph. About 50% of subjects sort mistakenly).

Men and women were equally productive in three respects. First, there is no statistical

[7] We used the short version of the NEO PI-R test in our experiment. The Danish NEO-PI-R Short Version consists of five 12-item scales measuring each domain. The Danish short version strongly correlates with the long test in such a way that Costa and McCrae (2003) conclude that if one only wanted to examine the Big Five factors, one could just as well use the short test instead of the full version.

difference between average output of men and women (59.58 vs. 61.14) according to a Wilcoxon rank sum test (WRST, $p = 0.483$) or a t-test ($p = 0.165$). Second, the overall distribution of output does not statistically differ by gender according to a Kolmogorov-Smirnov test (KS, $p = 0.131$). Finally, the share of subjects with output above or below 60 tasks does not differ by gender (57.4% for men and 59.1% for women, $p = 0.657$, WRST).

The only gender difference occurs in subjects with very low output. An output below 30 tasks is more common for men than women (6.6% vs. 2.0%, $p = 0.034$, WRST).[8] However, due to the large difference to the indifference point at 60 tasks, these subjects are unlikely to make sorting errors and are unlikely to have any relevance for our analysis. Indeed, we do not find sorting differences in subjects with output below 30 tasks compared to those with output between 30 and 60 tasks (32% vs. 26.5% Flat choices, $p = 0.667$, WRST).

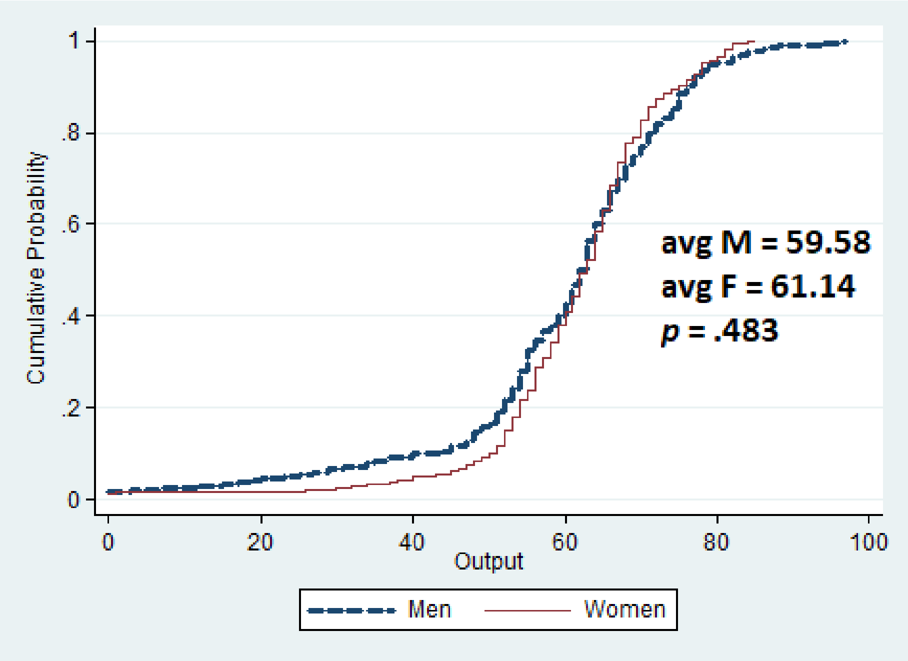


Figure 1 *Cumulative distribution of output (i.e. completed tasks) by gender; no significant differences except on the low end of the output spectrum. Legend shows averages for men and women; p-value for a WRST; N=715*

Figure 2 shows that highly productive women make more sorting mistakes than highly productive men. To see that this is so, we split the sample into those with output at or below

[8] As a consequence, we do observe heavier tails for the men's distribution than for the women's (kurtosis of 10.56 versus 5.95, respectively).

60 tasks (bars on the left, 299 subjects) and those with output above 60 tasks (416 subjects). The bars show the share of subjects choosing Flat. Sorting is far from perfect. If it were, the bars on the left would be at 100% and the bars on the right would be at zero. Instead, we find that sorting errors are common (about 50% of all subjects sort mistakenly), and that, overall, men choose Steep more often than women (73.5% vs. 66.2%, $p = 0.032$, WRST). Men with output below 60 tasks choose Flat significantly more often than men with output above 60 tasks ($p < 0.001$, WRST). Perhaps surprisingly, this comparison is only weakly significant for women ($p = 0.087$, WRST). Most importantly, well-performing women (i.e. women with output above 60) are much more likely to choose Flat than their male counterparts, see rightmost bars (18.9% vs. 30.2%, $p = 0.008$, WRST).

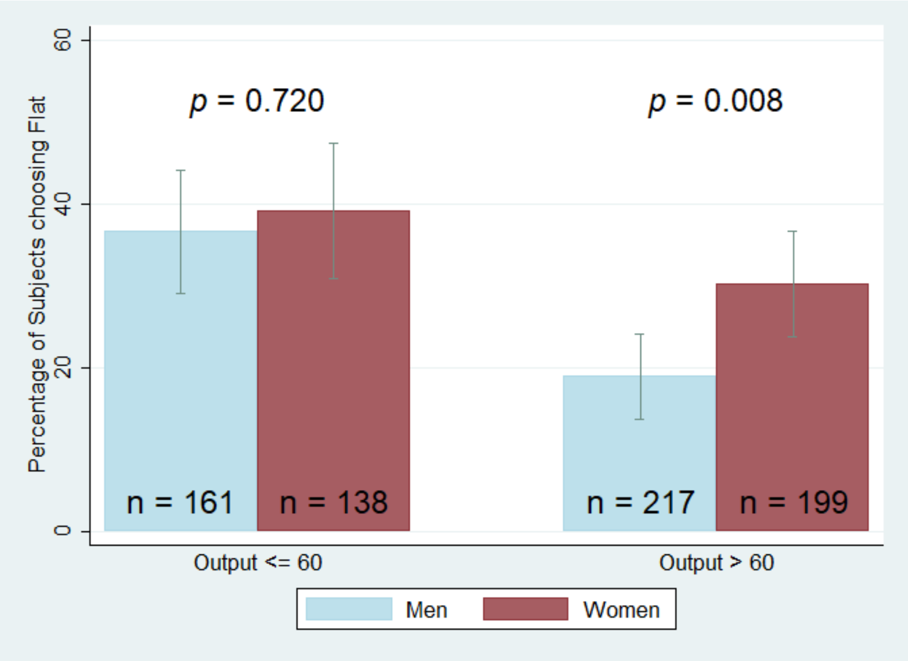


Figure 2 *Sorting into jobs by gender; women of high productivity (output $>$ 60 tasks) sort significantly more often into Flat than highly productive men; numbers above bars show p-values for a Wilcoxon rank sum test; N=715*

3.2 What explains sorting?

This section investigates mistaken beliefs about one's own output and two preference-driven reasons for why highly productive women sort into Flat at the detriment of their earnings: risk aversion and personality traits. It is important to investigate these latter explanations because they do not imply efficiency losses if they had any bite. The reason is that sorting into Flat would be the result of preference maximization and therefore not be a source of concern. In contrast, if sorting into Flat is driven by mistaken beliefs about one's own productivity (e.g. underconfidence), efficiency losses arise and policy interventions might therefore be in order (see conclusions for a discussion). We show below that beliefs are the main determinant for sorting choices.

An advantage of our design is that we control for many factors that might drive job selection other than beliefs. For example, we can rule out effects of strategic considerations and externalities (because both jobs involve individual-decision making tasks), and of preferences over jobs (because tasks are identical in the two jobs). We have already shown in section 2.2 that there are no differences in the behavioral response to the difference in marginal incentives to work across wage schemes. Hence, sorting choices are not explained by incentive effects. Despite all this, two important alternative explanations for gender sorting remain, namely risk preferences and personality. These factors are not controlled in our design, but we can test for their relevance by using elicited values.

Risk aversion is a plausible explanation for gender differences in sorting into Flat. Numerous previous studies have shown that women are more risk averse than men (see Eckel and Grossman 2008 or Croson and Gneezy 2009 for surveys), and this also holds in our sample according to a standard test in the multiple-price-list format developed by Holt and Laury (2002), ($p = 0.028$, WRST).[9] The reason why higher risk aversion in women may imply sorting into Flat is that the earnings variance is higher in Steep than Flat by design for given (uncertain) beliefs on output. In short, choosing Steep is more risky, and because women are more averse to risk, they may choose Flat more often than men for given average productivity beliefs above 60 tasks.

Another plausible explanation for gender sorting are personality traits, as overconfidence

[9] We use the number of safe choices as a measure of subjects' risk attitudes. See Csermely and Rabas (2016) for a comparative evaluation of alternative risk elicitation methods.

also appears to depend on general traits of an individual (Schaefer et al. 2004), especially emotional stability (also called neuroticism, see Müller and Schwieren 2012). People with a low value in emotional stability react adversely to negative experiences, so they might make safer choices to minimize the distress of choosing the wrong payment scheme. Women in our sample have a significantly lower score of emotional stability than men ($p < 0.001$, WRST) which is a common finding in the literature (see e.g. Schmitt et al., 2008).[10]

A Beliefs explain choices

Figure 3 shows that men are more optimistic about their performance than women on average.[11] They think they will complete 79.86 vs. 76.89 tasks for women, $p = 0.007$, WRST). However, we also find that the level of beliefs is generally too high. In fact, most subjects (68%) overestimate their output and subjects on average substantially overestimate their output which in fact was around 60 tasks for men and women alike (see Figure 2).[12] A plausible explanation for this substantial overestimation of output is that subjects were experienced (in the trial phase) only with simple tasks with few yellow cells when they made the job choice, and may have underestimated how much more difficult the task gets over time. While beliefs were generally overoptimistic, we also find that women hold beliefs that are more accurate than men's, but this might be an artifact of our design that tended to produce overoptimistic beliefs. To illustrate, suppose that women generally have lower beliefs than men (as is the case), and that beliefs are below the cutoff point (i.e. below 60 tasks, which is not the case). Then, men's beliefs would be more accurate.[13]

[10] We also find that men are less agreeable and open than women (see Table A2 for details). However, we could not find hypotheses in the literature suggesting how differences in these traits may translate to job sorting in our experiment.

[11] In the literature, a distinction is sometimes made between overconfidence and optimism, where optimism occurs when individuals, independent of their own performance, overestimate the probability of a success. Overconfidence occurs when individuals believe they perform better than they actually do. We use these two expressions interchangeably because the probability of a success (i.e. receiving a high wage) only depends on one's own performance in our setting.

[12] As we elicited two measures for productivity beliefs, one for each job, we use the mean of these two measures as our belief measure. This is unproblematic, as beliefs in both schemes are the same for over 80% of subjects, and we have established that there is no incentive effect (see sections 2.2 and A.2).

[13] Nonetheless, the finding that subjects generally overestimate their abilities and that women's beliefs are more realistic than men's is also found in Bordalo et al. (2017).

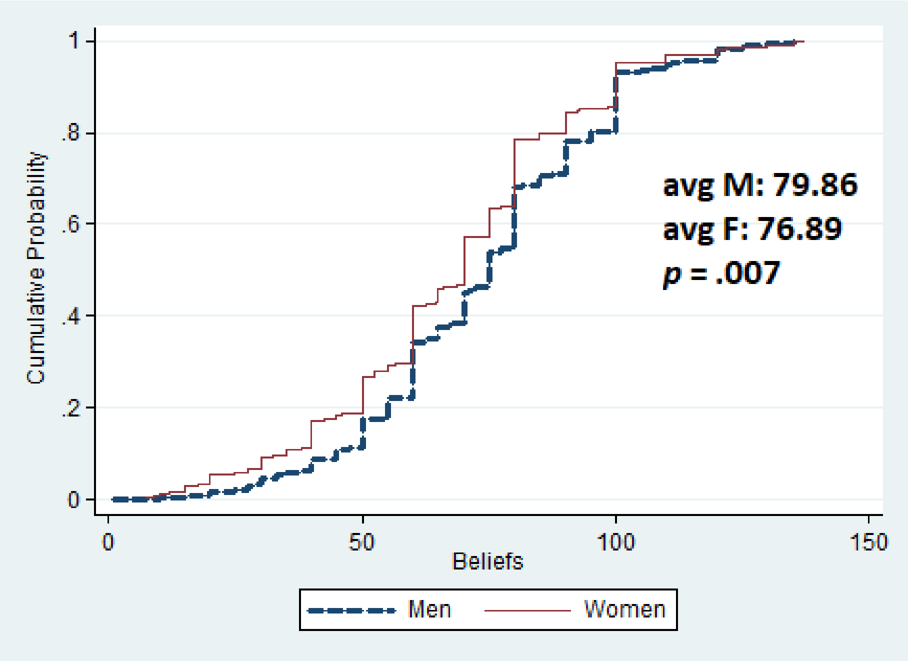


Figure 3 *Cumulative distribution plot for beliefs by gender; women have significantly lower beliefs; subjects with beliefs higher than 150 omitted for readability; averages for men and women on the right;* p*-value for a WRST; N=672*

Table 1 shows that beliefs are the main determinants of job sorting. The table shows probit regressions with the dependent variable being 1 if a subject chose Flat and 0 otherwise. Column 1 shows that, for all subjects, beliefs are a highly significant predictor for choice, meaning that more optimistic workers are more likely to choose Steep. One standard deviation increase in optimism (i.e. high beliefs) means that the probability to choose Flat falls by 19.2 percent. IQ is significant as well, meaning that more intelligent workers are less likely to choose Flat (one standard deviation increase in IQ reduces the probability to choose Flat by 5.3 percent). However, one needs to be careful with this interpretation since IQ to some extent proxies for productivity. In fact, IQ is correlated with output ($\rho = 0.2$, $p < 0.001$). In any case, there is no observable gender difference in IQ (see appendix Table A2 for details).

Controls include education, age and the Big5 personality measures, none of which are significant, and neither is out measure of risk aversion, as can be seen in row 2. The coefficient for female is weakly significant which indicates that there is a residual gender effect that is not captured by productivity beliefs, risk aversion or the controls (education, age, and the Big5 personality measures).

Columns 2 and 3 therefore present separate probit regressions for men and women. Results show that beliefs are highly significant when sorting is estimated separately by gender. The variable IQ continues to be significant for women but ceases to be significant for men, indicating that IQ seems to map into choices differently for men and women.

We use a dominance analysis as a robustness check for the results in Table 1. A dominance analysis determines the relative importance of independent variables in an estimation model based on their relative contribution to an overall model fit statistic. In our case, the general model is the probit regression used in Table 1, and the model fit statistic is the pseudo R^2. The dominance analysis reveals that beliefs have the highest relative importance, completely dominating all other variables, followed by IQ, overall as well as for men and women separately. Across all subjects, beliefs explain 75% of explained variance (see appendix Table A7 for details).

Table 2 addresses the concern that Table 1 may misrepresent the relevance of determinants like risk aversion and personality by giving all observations equal weight. When a subject's productivity belief is near 60 tasks, i.e. when the earnings lost due to sorting into the wrong job are low, other factors like risk attitudes or personality traits are likely to become a stronger predictor for choice. For example, risk aversion may induce a fully rational decision maker to choose Flat at a belief of 62 tasks but not at one of 82 tasks. We address this concern by repeating the regressions in Table 1 with higher weights for subjects with beliefs closer to 60 tasks.[14] Table 2 shows that in addition to beliefs - which remain strongly significant predictors of sorting (see row 1) - many other factors also significantly influence sorting with weighted observations. In particular, we see that risk attitudes are significant[15] for women as well as education, IQ is now also significant for men, and all personality measures turn out significant (last five rows), including emotional stability which seems to have differential effects for men and women. We are careful in interpreting the effects of personality traits on choice since we had no clear ex ante hypotheses (apart from emotional stability) on how they should influence sorting in this environment.

[14] The exact weight used is 60 minus the distance of the beliefs to 60, with a minimum weight of 1, i.e. $max(1, 60 - |60 - belief|)$.

[15] Our finding that risk aversion matters for those with output around 60 is reminiscent of the finding in Dohmen and Falk (2011) that risk attitudes play a bigger role in the piece-rate treatment for subjects who are marginal, i.e. who either contemplate longer or are close to the productivity threshold that makes them indifferent between fixed wages and piece rates.

Table 1: *Determinants of Gender Sorting*

	1 if Flat		
	all	women	men
Beliefs	$-.004^{***}$	$-.003^{***}$	$-.008^{***}$
	$(.000)$	$(.000)$	$(.000)$
Risk	$.010$	$.018$	$.001$
	$(.006)$	$(.008)$	$(.008)$
IQ	$-.016^{***}$	$-.019^{**}$	$-.011$
	$(.006)$	$(.009)$	$(.008)$
Female	$.054^{*}$		
	$(.032)$		
Controls	Yes	Yes	Yes
Pseudo R^2	$.13$	$.10$	$.20$
NObs	715	337	378

*Notes: marginal effects of a probit regression displayed; dependent variable is 1 (0) if Flat (Steep) was chosen; standard deviation below the coefficients in parentheses; belief is the ex ante productivity belief of a subject; risk means the number of safe choices in a Holt/Laury test; IQ is the number of correct answers for 20 of Raven's progressive matrices; controls include high education, age, and the Big5 personality measures, none of which are significant; stars are given as follows: ***:* $p < .01$, ***:* $p < .05$, **:* $p < .1$

Note that the pseudo R^2 is about the same overall in Table 2 as in Table 1, and even lower for men, despite inclusion of many additional significant coefficients. The reason is that behavior becomes more noisy as the costs of errors are lower for beliefs close to the indifference point, i.e. the variance in the dependent variable is higher. Repeating the dominance analysis for Table 2 reveals that beliefs have again the highest relative importance in this weighted regression, completely dominating all other variables, and this is true overall as well as for men and women separately. Beliefs account for about 70% of explained variance in the weighted regression (see appendix Table A8 for details). In conclusion, we find that (misaligned) beliefs are by far the main determinant of (mistaken) sorting.

Table 2: *Determinants of Gender Sorting - Weighted Beliefs*

	1 if Flat		
	all	women	men
Beliefs	$-.026^{***}$	$-.023^{***}$	$-.032^{***}$
	$(.001)$	$(.001)$	$(.001)$
Risk	$.023^{***}$	$.055^{***}$	$-.001$
	$(.003)$	$(.004)$	$(.005)$
IQ	$-.055^{***}$	$-.068^{***}$	$-.033^{***}$
	$(.003)$	$(.005)$	$(.004)$
Female	$.103^{***}$		
	$(.018)$		
High Education	$.036^{**}$	$.089^{***}$	$-.008$
	$(.018)$	$(.026)$	$(.025)$
Agreeableness	$-.000$	$.012^{***}$	$-.009^{***}$
	$(.002)$	$(.003)$	$(.002)$
Conscientiousness	$.006^{***}$	$.025^{***}$	$-.010^{***}$
	$(.002)$	$(.003)$	$(.002)$
Extraversion	$-.015^{***}$	$-.005^{**}$	$-.029^{***}$
	$(.002)$	$(.002)$	$(.002)$
Emotional Stability	$-.011^{***}$	$-.033^{***}$	$.014^{***}$
	$(.002)$	$(.002)$	$(.002)$
Openness	$-.012^{***}$	$-.035^{***}$	$.007^{***}$
	$(.006)$	$(.002)$	$(.002)$
Pseudo R^2	$.13$	$.15$	$.14$
NObs	715	337	378

Notes: This table repeats the regressions in Table 1, but observations are weighted according to the distance of a subject's belief to 60 with a minimum of 1 , i.e. $max(1, 60 - |60 - belief|)$*; marginal effects of a probit regression displayed; dependent variable is 1 (0) if Flat (Steep) was chosen; standard deviation below the coefficients in parentheses; risk means the number of safe choices in a Holt/Laury test; IQ is the number of correct answers for 20 of Raven's progressive matrices; high education means Bachelor degree or higher; personality measures come from the Big5; ***:* $p < .01$, ***:* $p < .05$, **:* $p < .1$

B Inconsistent choices just add noise

A possible concern with our claim from the previous section that mistaken beliefs drive job sorting is that sorting may be driven by inconsistent choices. Inconsistent job choices are not uncommon in out sample since only 79.7% of subjects are "consistent" in the sense that they choose jobs according to their beliefs.[16] However, this section argues that the job choices that are inconsistent with productivity beliefs do not explain that highly productive women sort into Flat but merely add noise.

Imagine the following situation: A man and a woman both have productivity beliefs of 70. As beliefs are correlated with output ($\rho = .072$, $p = 0.055$), both will also tend to have high output. Then, as men are more consistent than women among those with beliefs larger than 60 (87.1% vs. 80.9%, $p = 0.068$, WRST), the man might choose Steep (i.e. choose consistently), and the woman Flat (i.e. inconsistently). Therefore, differences in consistency could explain the gender sorting effect that highly productive women choose Flat too often.

But the concern is unwarranted because inconsistent subjects do not sort differently across genders: When output is below 60 tasks, 50% of men and 55% of women choose flat ($p = 0.675$, WRST), and when output is above 60 tasks, 53.1% of men and 41.5% of women choose flat ($p = 0.325$, WRST). Hence, if anything, highly productive women who are inconsistent choose Steep more often than men.[17]

Inconsistent people therefore only add noise to our data, as gender sorting is not explained by inconsistent choices. To provide a sharper image of how beliefs map into choices we only consider consistent subjects from here on and omit 145 inconsistent subjects from future analyses. However, we provide all analyses presented below using all subjects, including inconsistent ones, in appendix A.4 to show that our results do not change qualitatively when including inconsistent subjects.

[16] Subjects are defined as consistent if they choose Flat if their beliefs are lower than 60 and Steep if they are higher than 60. Note that subjects with beliefs of exactly 60 can never be inconsistent.

[17] For a complete breakdown of gender choices, beliefs and consistency see Section A.3 in the appendix.

C Highly productive women are underconfident

Table 3 classifies consistent subjects according to their beliefs (above vs. below a belief of 60 tasks, i.e. optimistic vs. pessimistic) and output (above vs. below 60 tasks, i.e. highly productive vs. unproductive). Optimistic but unproductive workers are overconfident in their abilities and make a type I sorting error while pessimistic but highly productive workers are underconfident and make a type II sorting error. The p-values indicate whether the number of men is different from the number of women in a cell according to a Wilcoxon Rank Sum test. The only significant difference occurs in the bottom right cell, i.e. we find that there are significantly more women who make a type II sorting error than men. In fact, 14.0% of women and 5.6% of men make a type II error ($p = 0.004$, WRST[18]), i.e. the type II error is more than twice as likely for women than for men. In other words, the combination of high productivity but low self-confidence (or pessimism) is much more common among women than men.

Table 3: *Classification by confidence and productivity*

	Output	
	< 60	> 60
Beliefs $>$ 60	Type I Error	Correct
Choose Steep	m=68, f=51	m=152, f=107
Optimistic	$p = .613$	$p = .115$
Beliefs $<$ 60	Correct	Type II Error
Choose Flat	m=31, f=21	m=15, f=29
Pessimistic	$p = .492$	$p = .004$

Notes: Classification of consistent subjects (N = 474) by beliefs and output, i.e. all subjects with beliefs above 60 tasks chose Steep and vice versa; "correct means that subjects have beliefs in line with output (subjects are neither over- nor underconfident in relation to the indifference point of 60 tasks); note that subjects who have output or beliefs of exactly 60 cannot be incorrect in their choice, so these 96 subjects are omitted; p-values for a WRST that tests the frequency of men and women in a given category; There are significantly more women in the bottom right cell

[18] This WRST tests a dummy variable that has value 1 if a consistent subject falls into the bottom right cell and 0 otherwise by gender. Subjects with output or beliefs of exactly 60 are excluded.

Figure 4 shows the frequency of correct vs. incorrect choices by beliefs and gender (M = male and F = female). We again exclude inconsistent subjects which implies that all subjects with beliefs below 60 tasks (those shown in left half of the figure) chose Flat and those with beliefs above 60 tasks (those in the right half of the figure) chose Steep. The labels "correct" and "incorrect" in the figure refer to whether beliefs are in line with output. For example, choices labeled as "incorrect" at beliefs below 60 tasks mean that subjects chose Flat but were of high productivity (i.e. output above 60 tasks), and "incorrect at beliefs above 60 tasks means that subjects chose Steep but were in fact of low productivity.

Five conclusions emerge from Figure 4. First, many more subjects choose Steep rather than Flat (bars in the right half are higher than in the left half). This is so because beliefs were generally highly optimistic, i.e. workers tended to have expectations above 60 tasks, and chose Steep in line with their beliefs. Second, mistaken sorting is more common among those who chose Steep than those who choose Flat. This can be seen by the fact that red segments of the bars are larger to the right than the left. The reason is that beliefs tend to be excessively optimistic , i.e. exceed output on average (by 23.67 for men vs. 19.79 for women, $p = 0.018$, WRST). Third, low productivity beliefs are more common among women than men, and vice versa for high beliefs. This finding is reflected in the figure by the fact that bars for women (marked with F) are higher than those for men in the left half of the figure, and vice versa in the right half ($p = 0.045$, WRST). Fourth, incorrect choices are more frequent for pessimistic women than for pessimistic men ($p = 0.010$, WRST), but this does not hold for optimistic men vs. women ($p = 0.591$, WRST). This significant difference illustrates our main result from Table 3, i.e. that highly productive but underconfident women are more likely than similar men to choose Flat. Fifth, incorrect choices seem to be more common when productivity beliefs are close to 60 tasks. This fact can be seen in the figure by noting that red portions of the bars tend to be larger in the center than at the borders of the figure. This result is in line with the fact established in Table 2 that when beliefs are near 60, factors other than beliefs matter for sorting.

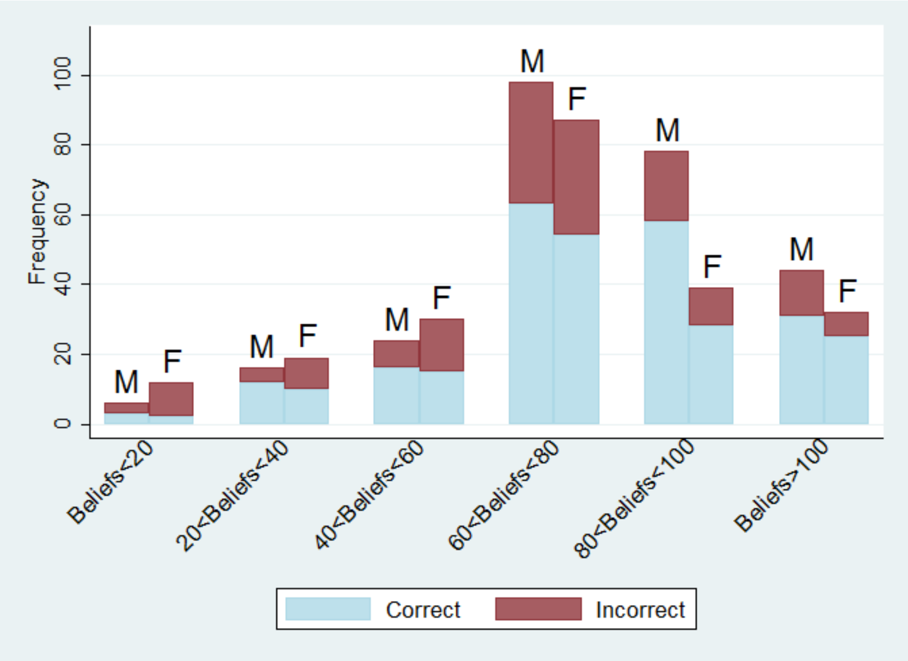


Figure 4 *Absolute number of correct choices by beliefs and gender; consistent subjects only; subjects shown in left half of the figure chose Flat, those in right half chose Steep; "Beliefs" indicates expected number of completed tasks; "Correct" means subjects have beliefs in line with true output, therefore "incorrect" indicates a sorting mistake; there are significantly more women than men with beliefs below 60 tasks; more incorrect choices for women with beliefs below 60 tasks; most incorrect choices overall when beliefs are near 60; N = 474.*

D What explains choices? - Summary

To sum up, we have established that the task is gender-neutral, i.e. there are no differences in output between men and women, that highly productive women sort significantly more often into Flat than similar men, and that women have lower beliefs about own output than men. We also show that risk attitudes or personality measures do on average not influence the job choice (they only have an effect on choices when beliefs are near 60 tasks, and in these cases sorting errors only tend to have a small effect on earnings), and the best predictor for sorting choices are the beliefs. We find that workers generally tend to overconfident but men are generally more overconfident than women. From these observations it follows that overconfidence-induced mistaken sorting into Steep is more common among men, and underconfidence-induced mistaken sorting into Flat is more prevalent for women. As the presence of inconsistent subjects cannot explain gender sorting (inconsistent women do not choose Flat more often), we conclude that differences in beliefs are the main explanation for sorting differences between men and women: Highly productive women are more than twice as likely to have low beliefs than men, and this fact explains why highly productive women tend to sort into inadequate (i.e. low-paying) jobs.

3.3 Earnings

To maximize their earnings, highly productive workers (output above 60 tasks) should sort into Steep and the unproductive workers should sort into Flat. Hence, sorting mistakes cause lost earnings and the size of losses increases linearly with the distance to the indifference point at 60 tasks.

Our main finding from the previous section that highly productive women mistakenly sort into Flat more often than similar men implies that these women also earn less and we indeed find that this earnings difference is significant ($p = 0.005$, t-test). However, we do not find that (consistent) women generally earn less than men (91.28 DKK for men and 91.85 DKK for women on average, $p = 0.708$, t-test).[19] The reason is that type I sorting errors (being low-productive but choosing Steep) tend to be more costly for men than women. While there are no gender differences in the frequency of type I errors ($p = 0.613$, WRST), there are more men with very low outputs (see Figure 1), so type I errors are more costly for men, leading to significantly lower earnings for unproductive men than women ($p = 0.046$, t-test).

We conclude that men and women make different mistakes in job sorting. Men tend to sort too much into Steep, and women too much into Flat, and these mistakes also translate into lost earnings. This means that low-productive men and highly productive women do not realize their earnings potential. Men fail to do so because they tend to be overconfident, women because they are underconfident, i.e. women tend to be overly pessimistic about their ability to produce. However, because the two types of mistakes tend to cancel out, we find no significant overall effect of gender sorting on earnings.

[19] There is also no difference if we include inconsistent subjects; see section A.4 in the appendix.

4 External validity: Generalization to the labor market

This section shows that the earnings losses we find for underconfident women in the virtual laboratory translate to the field. More precisely, we find that highly productive women who have pessimistic beliefs about their own productivity and therefore sort into low-paying jobs in the lab also earn less than highly productive but optimistic women in the field.

We are able to show that this is so by matching our experimental data to income data from official registers (see section 2.3 for details). In the analysis below, we focus on annual taxable incomes. This measure includes all taxable (gross) income from wages, pensions, fees etc., as well as income from entrepreneurial activity, but not capital income or income from abroad (like investment income).[20] One might be concerned with the accuracy of tax data due to tax evasion and black market activity, especially if these activities are systematically different across gender or productivity levels. We can of course not exclude such correlations but point out that Danish tax reporting is among the most complete in the world. According to Kleven (2014), Denmark has "very wide coverage of third-party information reporting and more generally, well-developed information trails that ensure a low level of tax evasion". They also have broad tax bases that "encourage low levels of tax avoidance.

Table 4 parallels Table 3 in structure and shows how output and beliefs in the experiment relate to participants' total annual earnings in the field. As we do in previous sections, we again focus on consistent subjects, i.e. those who sort into jobs according to their beliefs in the lab. However, the results reported in this section also hold for all subjects (see appendix A.4). The upper number in each cell of Table 4 shows income in 10,000 DKK, the numbers in parentheses show the number of participants.

Three observations are in order. First, in our sample, men earn significantly more than women on average (303,410 DKK or about 40,500 Euro vs. 246,320 DKK or about 32,800 Euro, $p < 0.001$, t-test). Many factors other than sorting might explain this gender difference in income, including higher incidence of part-time employment among women (working hours per

[20] See http://www.dst.dk/da/Statistik/dokumentation/Times/ida-databasen/ida-personer/slon.aspx for a detailed description (Danish only).

week are 38.01 for men and 34.98 for women, $p < 0.001$, WRST), differences in education[21] and experience, public sector employment (18.1% of men and 22.2% of women, $p = .041$, WRST) or labor market discrimination. Unfortunately, our study must remain silent on the relevance of these other factors or their relative importance in explaining the gender wage gap.

Second, optimistic workers (i.e. those with beliefs above 60 tasks) earn more than pessimistic ones ($p < 0.001$, t-test). This can be verified in Table 4 by noting that all income numbers in the top row are higher than the numbers in the bottom row.Strikingly, our experimental measure of confidence in a simple counting task that admittedly has little resemblance to the work tasks in many occupations seems to predict incomes in the field. While we think that this is a remarkable finding, we hasten to say that these are of course just correlations and do not allow us to infer causality. It may, for example, well be that subjects hold high beliefs about their output in the virtual lab because they work in high-income jobs and are therefore confident about their abilities in general.

Third, subjects with output above 60 tasks do not have higher incomes than those with low output ($p = 0.475$, t-test).

We cannot find any significant differences for men across cells in Table 4. We think that this is due to high earnings volatility for men. In fact, the standard deviation of incomes is 268,190 DKK for men but only 163,116 DKK for women ($p < 0.001$, Levene-test).

However, we do find two significant differences across cells for women. One concerns differences among women who sort correctly. That is, earnings of women who are optimistic and productive (top right) are higher than of women who are pessimistic and unproductive (bottom left cell, $p < 0.001$, t-test). The most important difference in the light of our discussion above is that we find that optimism vs. pessimism makes a significant difference for highly productive women. Earnings for women in the top right cell are almost twice as high than for women in the bottom

[21] Education is an unlikely explanation in Denmark. Secondary education levels are very similar for men and women (69.3% for men and 71.3% for women), and women even have a higher frequency of tertiary education than men (24.4% for men and 32.9% for women). Furthermore, we detect no differences in higher education in our sample (see Table A2). But as is the case across the EU-27, selection into different fields of study is observable also in Denmark. Source: Gender Equality in Denmark - Country Profile - European Parliament - 2013

right cell (312,600 DKK vs. 171,000 DK, $p < 0.001$, t-test).[22]

Table 4: *Mean taxalbe income by productivity and optimism*

	Output < 60		Output > 60	
	Men	Women	Men	Women
Beliefs > 60	36.06	22.24	31.28	31.26
	(68)	(51)	(152)	(107)
Beliefs < 60	26.94	17.80	21.87	17.10
	(31)	(21)	(15)	(29)

Notes: Mean income is the gross sum of all incomes per year for our subjects in 10,000 DKK. This measure includes all taxable income from wages, pensions, fees etc., as well as income from entrepreneurial activity, but not capital income, income from abroad or tax-free income; women of high output earn significantly more when they have high beliefs; inconsistent subjects and subjects with beliefs or output of exactly 60 omitted; number of observations in parentheses; N=474

Table 5 confirms that the result that underconfidence is costly for highly productive women in the labor market also holds in a regression analysis. The dependent variable is the log of the sum of gross income from tax reports defined as before, and the independent variables are output in number of completed tasks in the experiment, a dummy for beliefs below 60 tasks, IQ, education, age, age squared and risk attitudes.

Column 4 in Table 5 shows the effect of optimism vs. pessimism for highly productive women when controlling for productivity in the experiment and a host of other factors. The coefficient in the first line shows that highly productive but pessimistic women earn 41.9% ($=e^{-0.870}$) less on average than otherwise identical but optimistic women, and vice versa. This finding shows that the mistaken sorting we have identified in the virtual lab (see Figure 2) translates to taxable income in the field.

Column 3 shows the analogous result for men, but the corresponding coefficient (see row 1) is only weakly significant and only half the size of the effect for women. This means that

[22] We investigated whether there are notable differences in the kind of occupation, working hours or other parameters related to job choice between highly productive optimistic vs. pessimistic women. But we were unfortunately not able to identify any such differences due to the low number of subjects in the bottom right cell.

men are less punished than women for being underconfident in terms of incomes earned in the labor market.[23] Column 1 shows a weakly significant negative coefficient in row 1. This seems to indicate that men are in fact rewarded for being overconfident when they are of low productivity. The negative coefficient in row 1 indicates that unproductive men who chose correctly (beliefs and output below 60) earn a staggering 61.4% less than unproductive men who make are overconfident (beliefs above 60, but output below 60 tasks) if we want to believe the estimate (note that it is only weakly significant).

We are not surprised that output is not significant in any of the regressions in row 2, as income was not significantly different for subjects with high vs. low output in Table 4. However, we are surprised to see that IQ is not significant either in row 3 in any of the regressions, as IQ predicts lifetime income rather well in the literature (see e.g. Zagorsky 2007). Even if we run a regression on income with only IQ as the independent variable (not reported), it is not significant in all 4 columns, indicating that our particular measure of IQ might not be well-suited for the purposes of analyzing income. High education, age and age^2 turn up significant in most of the columns as we would expect, but not in all of them.[24] Especially column 1 is puzzling (pessimistic and unproductive men), as age does not seem to have an influence on earnings in this group. These men seem to be an especially heterogeneous group, as the R^2 is much lower than in the other columns. Finally, risk in row 7 is not significant in any column in line with our findings in section 3.2.A.[25]

Of course, the evidence provided here is only correlational as we cannot claim a causal link between the results in our experiment and income in the labor market. We conclude that women of high ability but low confidence in their productivity earn significantly less on the labor market than their equally productive peers (this result also holds for all subjects, including the inconsistent ones, see appendix section A.4, Table A9).

[23] However, we do not find significant differences in the two rightmost coefficients in row 1, $p = 0.223$, Chow-test.

[24] According to our estimates, income increases for low-productive women up to an age of 53.5 years, and for highly productive up to an age of 58.7 years, and falls thereafter. For highly productive men, it increases until age 55.9. We speculate that age is not significant in column1 because the group is more heterogeneous than the others, as is suggested by the low R^2.

[25] Nonetheless, risk attitudes should not be disregarded as an influence on sorting behavior and incomes. For example, Bonin et al. (2007) show that individuals with low willingness to take risks are more likely to sort into occupations with low earnings risk in the German labor market.

Table 5: *Determinants of taxable income (sum of gross incomes)*

	Output< 60		Output> 60	
	Men	Women	Men	Women
Beliefs < 60	$-.488^{*}$	$.115$	$-.432^{*}$	$-.870^{***}$
	$(.278)$	$(.280)$	$(.253)$	$(.206)$
Output	$-.011$	$-.002$	$-.008$	$.027$
	$(.008)$	$(.010)$	$(.010)$	$(.015)$
IQ	$-.029$	$-.065$	$.037$	$-.015$
	$(.042)$	$(.053)$	$(.028)$	$(.031)$
High Education	$.626^{**}$	$.668^{**}$	$.198$	$.041$
	$(.283)$	$(.306)$	$(.160)$	$(.201)$
Age	$.068$	$.170^{***}$	$.199^{***}$	$.108^{***}$
	$(.046)$	$(.046)$	$(.030)$	$(.039)$
Age$^2/100$	$-.055$	$-.159^{***}$	$-.178^{***}$	$-.092^{**}$
	$(.046)$	$(.046)$	$(.032)$	$(.042)$
Risk	$-.050$	$-.009$	$.010$	$.032$
	$(.051)$	$(.042)$	$(.027)$	$(.030)$
Constant	1.662	-1.379	-1.949	-1.553
	(1.257)	(1.287)	$(.994)$	(1.446)
Adj. R^2	$.119$	$.247$	$.321$	$.213$
NObs	97	69	161	134

*Notes: OLS regressions; dependent variable is the log of annual gross income from tax reports in DDK divided by 10,000; highly productive women with high beliefs have higher incomes; "Beliefs $<$ 60" is a dummy variable; "High Education" takes the value 1 if subject has a Bachelor degree or higher; "Risk" means the number of safe choices in a Holt and Laury test; standard deviation below the coefficients in parentheses; ***: $p < .01$, **: $p < .05$, *: $p < .1$*

5 Conclusion

This paper has used an internet experiment to shed new light on how beliefs about one's own productivity shape gender sorting into jobs. We use a gender-neutral work task and let 715 participants from all walks of life in the Danish population choose between two jobs under standard procedures and protocols for a laboratory experiment. One of these jobs is more profitable for the high productive and the other for the low productive. Our design is chosen such that beliefs about one's own productivity, i.e. self-confidence, rather than strategic concerns, concerns of how others are affected by one's work, or a (dis-)taste for competition determine sorting.

We find significant gender differences in the sense that women of high ability tend to be underconfident and therefore sort into a job profitable for low-productive workers (i.e. they make a type II sorting error). In contrast, men of low ability tend to be overconfident and therefore sort into jobs profitable for highly productive workers (i.e. they make a type I sorting error). This tendency is particularly pronounced among the highly productive women. In fact, highly productive women are more than twice as likely to have pessimistic beliefs than men, and this fact explains why highly productive women tend to sort into inadequate (i.e. unprofitable) jobs. Hence, highly productive women earn less than they could have earned because they sort into the "wrong type of job, and they do so because they expect to perform less well in the work task than they effectively do.

We match the experimental data with high-quality income data from official (tax) registers with the help of Denmark's statistics agency. We find that that highly productive women who sorted into unprofitable jobs because of underconfidence in the experiment also earn significantly less than equally productive women who are not underconfident in the field. Hence, the underconfidence effect on income we find in the lab seems to have remarkable external validity.

It is remarkable that we find these significant differences in Denmark which is by many measures one of the countries in the world with the highest degree of gender equality. Since the prevalence of underconfident women is in our view largely a cultural phenomenon, shaped by role models in the family and society and in particular in education, we suspect that effects of underconfidence on labor market outcomes might even be stronger in other countries.

It is worth noting that a free labor market generally has the tendency to correct sorting mistakes. If an individual makes a type I sorting error and works in a job for which he is underqualified, market forces could possibly remedy the situation, as employers might notice the inability of the worker and take appropriate measures. However, a type II sorting error is much less likely to be corrected, as an employer has no incentive to remove an overqualified, highly productive worker from his/her position. This, in itself, can explain why the effects of the experiment carry over to the labor market for women but not for men.

As far as policy implications are concerned, we need to understand whether people's job choices are efficient or not. On the one hand, if gender sorting were explained by differences in risk attitudes or personality, one might say that the workers act as utility maximizers in sorting into a job, and therefore no intervention is needed. But as far as we can gather from the information gained in our experiment, false sorting of women occurs in part due to false low beliefs about own productivity, which is inefficient and some of women's talents are lost. An intervention like quotas or affirmative action might be beneficial for efficiency reasons, if the women who are prone to type II sorting errors can be influenced by such interventions.[26] However, men who are prone to a type I sorting error in the experiment have similar or even higher earnings in the labor market as their peers, and more research is needed to gauge whether that is efficient or inefficient to the labor market as a whole.

[26] For more information on the subject of gender sorting, competitiveness and affirmative action, see Niederle et al. (2013).

References

Augusto De Araujo, F., Carbone, E., Conell-Price, L., Dunietz, M. W., Jaroszewicz, A., Landsman, R., Lam, D., Vesterlund, L., Wang, S. & Wilson, A. J. (2015). The effect of incentives on real effort: Evidence from the slider task. *CESifo Working Papers* No. 5372.

Azmat, G. & Petrongolo, B. (2014). Gender and the labor market: What have we learned from field and lab experiments?. *Labour Economics*, 30, 32-40.

Bayard, K., Hellerstein, J., Neumark, D. & Troske, K. (2003). New evidence on sex segregation and sex differences in wages from matched employee-employer data. *Journal of Labor Economics*, 21(4), 887-922.

Blau, F. D. & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and sources. *Journal of Economic Literature*. forthcoming.

Bonin, H., Dohmen, T., Falk, A., Huffman, D. & Sunde, U. (2007). Cross-sectional earnings risk and occupational sorting: The role of risk attitudes. *Labour Economics*, 14(6), 926-937.

Bordalo, P., Coffman, K. B., Gennaioli, N. & Shleifer, A. (2017). Beliefs about gender. *National Bureau of Economic Research*, No. w22972.

Brandts, J., Groenert, V. & Rott, C. (2015). The Impact of Advice on Women's and Men's Selection into Competition. *Management Science*, 61(5), 1018-1035.

Buser, T., Niederle, M. & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *Quarterly Journal of Economics*, 129(3), 1409-1447.

Cadsby, C. B., Song, F. & Tapon, F. (2007). Sorting and incentive effects of pay for performance: An experimental investigation. *Academy of Management Journal*, 50(2), 387-405.

Cappelen, A. W., Moene, K. O., Sørensen, E. Ø. & Tungodden, B. (2013). Needs versus entitlementsan international fairness experiment. *Journal of the European Economic Association*, 11(3), 574-598.

Card, D., Cardoso, A. R. & Kline, P. (2015). Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *Quarterly Journal of Economics*, 131(2), 633-686.

Cárdenas, J. C., Dreber, A., Von Essen, E. & Ranehill, E. (2012). Gender differences in competitiveness and risk taking: Comparing children in Colombia and Sweden. *Journal of Economic Behavior and Organization*, 83(1), 11-23.

Cason, T. N., Masters, W. A. & Sheremeta, R. M. (2010). Entry into winnertake-all and proportional-prize contests: An experimental study. *Journal of Public Economics*, 94(9), 604-611.

Corgnet, B., Hernán-González, R. & Rassenti, S. (2015). Peer Pressure and Moral Hazard in Teams: Experimental Evidence. *Review of Behavioral Economics*, 2(4), 379-403.

Cortes, P. & Pan, J. (2017): Occupation and Gender. *IZA DP* No. 10672

Croson, R. & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2), 448-474.

Csermely, T. & Rabas, A. (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.

Datta Gupta, N., Poulsen, A. & Villeval, M. C. (2013). Gender matching and competitiveness: Experimental evidence. *Economic Inquiry*, 51(1), 816-835.

Dohmen, T. & Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review*, 101(2), 556-590.

Dreber, A., von Essen, E. & Ranehill, E. (2011). Outrunning the gender gap - boys and girls compete equally. *Experimental Economics*, 14(4), 567-582.

Eckartz, K. (2014). *Task enjoyment and opportunity costs in the lab: The effect of financial incentives on performance in real effort tasks*. Jena Economic Research Papers No. 2014-005.

Eckartz, K., Kirchkamp, O. & Schunk, D. (2012). How do incentives affect creativity? *CESifo Working Paper Series*, No. 4049

Eckel, C. C. & Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results*, 1, 1061-1073.

Gneezy, U., Leonard, K. L. & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, 77(5), 1637-1664.

Gneezy, U., Niederle, M. & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, 118(3), 1049-1074.

Gneezy, U. & Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review Papers and Proceedings*, 92(2), 337-381.

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

Kamas, L. & Preston, A. (2012). The importance of being confident; gender, career choice, and willingness to compete. *Journal of Economic Behavior and Organization*, 83(1), 82-97.

Kleven, H. J. (2014). How can Scandinavians tax so much?. *Journal of Economic Perspectives*, 28(4), 77-98.

Larkin, I. & Leider, S. (2012). Incentive schemes, sorting, and behavioral biases of employees: Experimental evidence. *American Economic Journal: Microeconomics*, 4(2), 184-214.

Levanon, A. & Grusky, D. B. (2016). The Persistence of Extreme Gender Segregation in the Twenty-first Century 1. *American Journal of Sociology*, 122(2), 573-619.

Lewis, G. & Frank, S. (2002). Who wants to work for the government? *Public Administration Review*, 62(4), 395-404.

Lordan, G. & Pischke, J. S. (2016). Does Rosie like riveting? Male and female occupational choices. *National Bureau of Economic Research*, No. w22495.

McCrae, R. R. & Costa, P. T. (2003). *Personality in adulthood: A five-factor theory perspective*. Guilford Press.

Moore, D. A. & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, 115(2), 502.

Müller, J. & Schwieren, C. (2012). Can personality explain what is underlying women's unwillingness to compete? *Journal of Economic Psychology*, 33(3), 448-460.

Niederle, M. & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, 122(3), 1067-1101.

Niederle, M. & Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, 3(1), 601-630.

Niederle, M., Segal, C. & Vesterlund, L. (2013). How costly is diversity? Affirmative action in light of gender differences in competitiveness. *Management Science*, 59(1), 1-16.

Raven, J. (2003). Raven progressive matrices. In *Handbook of nonverbal assessment* (pp. 223-237). Springer US.

Reuben, E., Rey-Biel, P., Sapienza, P. & Zingales, L. (2012). The emergence of male leadership in competitive environments. *Journal of Economic Behavior and Organization*, 83(1), 111-117.

Reuben, E., Sapienza, P. & Zingales, L. (2015). Taste for competition and the gender gap among young business professionals. *National Bureau of Economic Research*, No. w21695.

Schaefer, P. S., Williams, C. C., Goodie, A. S. & Campbell, W. K. (2004). Overconfidence and the big five. *Journal of Research in Personality*, 38(5), 473-480.

Schmitt, D. P., Realo, A., Voracek, M. & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, 94(1), 168.

Sutter, M. & Rützler, D. (2010). Gender differences in competition emerge early in life. IZA Discussion Paper No. 5015

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

van Veldhuizen, R. (2017). *Gender differences in tournament choices: Risk preferences, overconfidence or competitiveness?*. WZB Discussion Paper No. SP II 2016-207.

Zagorsky, J. L. (2007). Do you have to be smart to be rich? The impact of IQ on wealth, income and financial distress. *Intelligence*, 35(5), 489-501.

A Appendix

A.1 Sample description

A Comparison Danish population - experiment

Table A1: *Representativeness of Sample*

	Experiment Fraction	Danish Pop. Fraction
Gender:		
Female	47.1%	50.2%
Age:		
18-30 years	16.2%	18.5%
30-44 years	25.9%	29.1%
45-59 years	38.5%	27.0%
60-80 years	21.2%	25.3%
Education (highest completed):		
Basic (up to 10 years)	8.5%	26.3%
High school or vocational	22.5%	45.4%
Medium tertiary education	50.1%	21.1%
Long tertiary education	18.9%	7.1%
Salary:		
Low (< Dkr. 300,000/year)	42.3%	65.9%
Medium (Dkr. 300,000-400,000/year)	28.4%	19.1%
high (> Dkr. 400,000/year)	29.3%	15.0%

Notes: n(experiment)=715, N(population)=4,457,659; for gender and age, the data in the column Danish population summarizes individuals between 18-80 years of age; for education the population is restricted to individuals between 20-69 (N=3,644,915)

B Descriptive Statistics

Table A2: *Background statistics and personality*

Variable	Description	Mean	Male	Female	p-value	Min	Max
Female	1 if female	0.47				0	1
Age	participant's age	47.60	48.23	47.00	.264	18	81
High Education	college degree or higher	0.53	0.52	0.54	.559	0	1
IQ	corr.answers on IQ test	9.03	9.08	8.98	.660	0	19
Agreeableness	Big 5 Factor	32.15	30.74	33.73	.001	12	46
Conscientousness	–	33.11	32.86	33.39	.196	9	47
Extraversion	–	30.62	30.89	30.33	.225	6	47
Neuroticism	–	19.30	17.64	21.15	.001	1	46
Openness	–	27.05	26.65	27.50	.065	8	47
Risk Aversion	Holt/Laury	4.23	4.01	4.47	.028	0	10

Notes: N=715; p-values for a t-test of mean difference between male and female

Table A3: *Experimental Variables*

Variable	Description	Mean	Male	Female	p-value
Scheme	1 if worked under Flat	0.41	0.40	0.42	
Choice	1 if Flat was chosen	0.30	0.26	0.34	.032
Output	correct answers	60.31	59.58	61.14	.165
Earnings	payment in DKK	90.67	90.14	91.27	.413
Absolute Belief, Steep	belief about output	80.40	82.10	78.56	.002
Absolute Belief, Flat	–	76.50	77.62	75.22	.004

Notes: N=715; p-values for a t-test of mean difference between male and female; no p-value for scheme as subjects were in part randomly assigned the payment schemes

A.2 Disentangling the Sorting and Incentive Effect

With our design, we are able to distinguish between the average sorting effect and the average incentive effect of choosing/being forced into a particular payment scheme. Consider the following definitions:

$$\bar{x^T} = \bar{x_{C,s}} - \bar{x_{C,f}} \tag{1}$$

$$\bar{x^T} = \bar{x^I} + \bar{x^S} \tag{2}$$

$$\bar{x^T} = \bar{x_s^I} + \bar{x_s^S} \tag{3}$$

$$\bar{x^T} = \bar{x_f^I} + \bar{x_f^S} \tag{4}$$

$$\bar{x_s^I} = \bar{x_{C,f}} - \bar{x_{F,s}} \tag{5}$$

$$\bar{x_s^S} = \bar{x_{C,s}} - \bar{x_{F,s}} \tag{6}$$

$$\bar{x_f^I} = \bar{x_{C,s}} - \bar{x_{F,f}} \tag{7}$$

$$\bar{x_f^S} = \bar{x_{C,f}} - \bar{x_{F,f}}, \tag{8}$$

where x stands for output, the superscripts $T/S/I$ stand for Total/Sorting/Incentive effects, the subscripts C/F stand for choice/forced groups, and the subscripts s/f stand for Steep/Flat wage schemes.

Equation (1) states that the average total effect $\bar{x^T}$ is the difference between the average output $\bar{x_{C,s}}$ of the group that chose Steep and the average output $\bar{x_{C,f}}$ of the group that chose Flat. Note that this is the effect that would also be observable and identifyable in the field.

Equation (2) states that the average total effect consists of two parts: The average incentive effect $\bar{x^I}$ and the average sorting effect $\bar{x^S}$. This is because when a person chooses to work under a different work compensation scheme, there are two effects that might explain any changes in his output: The different incentive system of the new scheme (incentive effect), and the fact that he himself chose such a scheme (sorting effect). In the field, we are not able to identify and disentangle these effects, as the counterfactuals are missing.

Equations (3) and (4) explain that this disentanglement of the total effect can be measured from two sides, either as the sum of the average incentive and sorting effects for the Steep group ($\bar{x_s^I}$ and $\bar{x_s^S}$), or as the sum of the average incentive and sorting effects of the Flat group ($\bar{x_f^I}$ and $\bar{x_f^S}$).

How these different incentive and sorting effects can be identified with our experiment can be seen in equations (5) and (6). The average incentive effect seen from the Steep side $\bar{x_s^I}$ is the difference in the average output of the choice forced group ($\bar{x_{C,f}}$) and the forced Steep group ($\bar{x_{F,s}}$), as these two groups chose the same wage scheme (Steep), but worked under different schemes. By comparing these two groups we can observe the counterfactual of subjects who chose the same wage scheme but worked under different ones, thereby identifying the incentive effect of working under different wage schemes.

The same is true of the average sorting effect $\bar{x_s^S}$ seen from the Steep side (Equation (6)): It is the difference of the average output of the choice Steep group ($\bar{x_{C,s}}$) and the average output of the forced Steep group ($\bar{x_{F,s}}$). By comparing these two groups we can observe the counterfactual of subjects who worked under the same wage scheme, but chose different wage schemes in the beginning, therefore identifying the sorting effect.

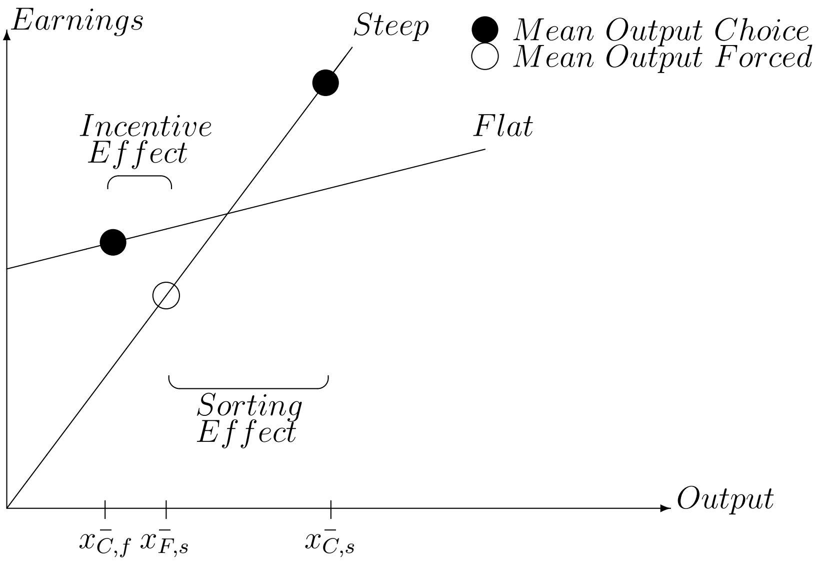
Analogously, the incentive and sorting effects from the flat side can be identified according to equations (7) and (8). The total effect should of course be the same, independent from whether you look at the decomposition from the steep or from the flat side.

The intuition behind these effects is the following: As the sorting effect measures differences in output between participants who chose Steep vs. Flat, we expect participants who chose Steep to have a higher output, as this choice maximizes earnings for high productivity workers. Also, as the incentive effect measures the difference in output of those who work under Steep vs. Flat, we normally would expect those in the steep scheme to have a higher output due to the higher marginal utility in the steep scheme (1.5 vs. 0.5 points). However, due to our particular parameter choices, we expect no incentive effect to be present because the marginal utility is positive in both wage schemes.

The total effect now effectively combines the sorting and incentive effects as it measures the output differences in the choice Steep vs. the choice Flat group. Note that if just the total effect were measured, it would not be transparent whether the difference in output could be attributed to a sorting or an incentive effect.

To illustrate the different effects, consider the following illustration. The black dot to the right represents the mean output of the choice Steep group, while the black dot to the left is the mean output of the choice Flat group. The white dot represents the mean output of the forced Steep group; the 4th group is omitted for illustration purposes. Note that this illustrates the decomposition of the total effect from the flat side, and that this graph is only an example of possible mean output levels.

Disentangling the Total Effect



The total effect can be measured by comparing the outputs of the choice Steep and the choice Flat groups (two black circles).

By comparing the average outputs of the forced Steep with the choice Flat group (the two left circles), we are able to get a measure of the incentive effect, as both groups chose the same scheme, but worked under different ones.

By comparing the choice Steep with the forced Steep group (the two right circles), we can capture the sorting effect, as they have chosen different wage schemes, but no incentive effect can be present because both groups worked under the same incentive scheme, i.e. they get the same amount of points per correct answer.

There is a Total Effect

We begin by establishing that there is in fact a Total effect present by comparing outputs of the CFlat vs. the CSteep group (MWU-Test, $p < 0.001$) - overall output is significantly higher for people who choose Steep over Flat and work under their preferred scheme (see Figure A1).

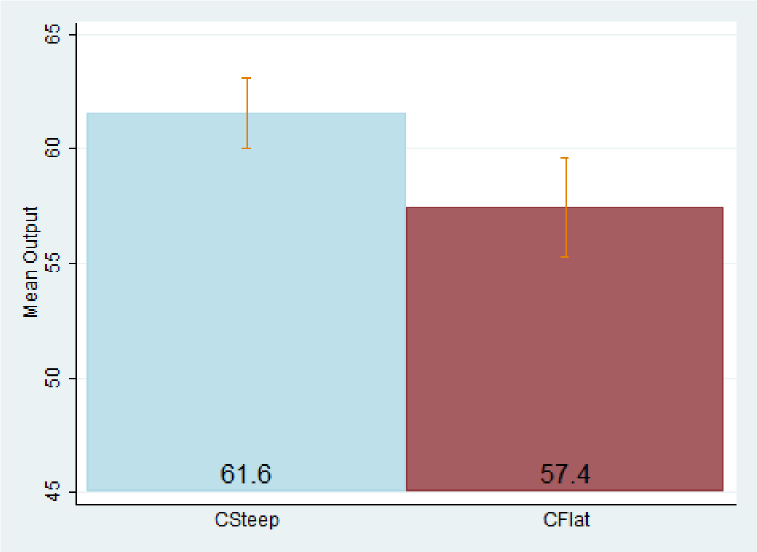


Figure A1 *Total effect is strong and evident between choice Steep and choice Flat groups; Difference in Mean Output*$=4.11$

There is no Incentive Effect

As was explained in the previous chapter, we can pin down the incentive effect by testing FSteep vs. CFlat and FFlat vs. CSteep; the p-values for the corresponding MWU-Tests are $p = 0.623$ and $p = 0.846$, respectively (see Figure A2). Furthermore, we also do not detect distributional differences ($p = 0.625$ for FSteep vs. CFlat and $p = 0.846$ for FFlat vs. CSteep; Krukal-Wallis test). We conclude that no incentive effect is present. This is reasonable because contrary to similar experiments, the piece rate (and therefore the marginal incentive to answer an additional question) is strictly positive. Furthermore, fatigue is unlikely to occur as the task ran for only 15 minutes, 93% of subjects worked the whole time (no difference in Steep vs. Flat, $p = 0.888$), and nearly noone took a break in between (again no difference between wage schemes, $p = 0.881$).

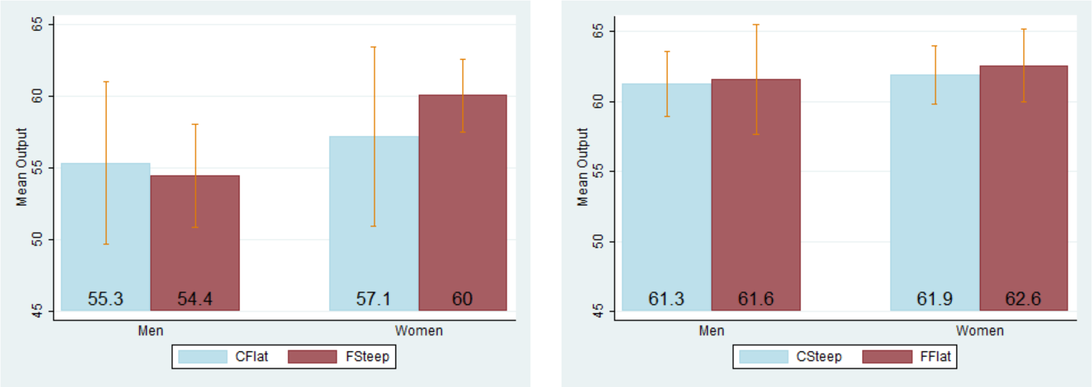


Figure A2 *No incentive effect, and also no incentive effect across genders; left graph compares CFlat vs. FSteep, right graph compares CSteep vs. FFlat; from left to right, conducting MWU-Tests for incentive effects, the p-values are:* $p = 0.959$, $p = 0.593$, $p = 0.841$, $p = 0.528$

As there exists a total effect but no incentive effect, we expect the sorting effect to be strong and evident. The corresponding MWU-Tests give p-values of $p < 0.001$ (see Figure A3). We conclude that the sorting effect is strong in our sample, and that the subjects worked as hard as they could in each group. Therefore, observed productivity is an unbiased measure of underlying productivity in this task.

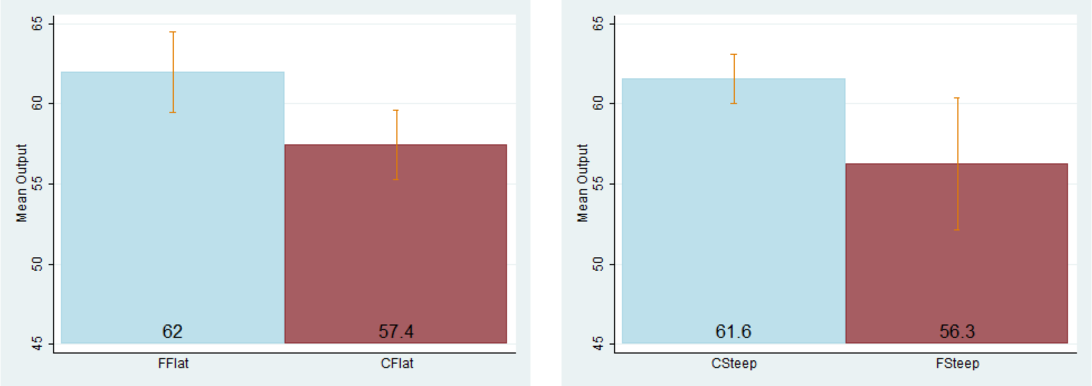


Figure A3 *Sorting Effect is significant; left graph compares CFlat vs. FFlat, right graph compares CSteep vs. FSteep*

A.3 Beliefs and Consistency

Tables A4, A5 and A6 give complete breakdowns of choices, output and consistency, overall as well as for men and women separately

Table A4: *Consistent Beliefs - all subjects*

subjects with output<60				subjects with output>60			
40% (280)				60% (416)			
choose Steep		choose Flat		choose Steep		choose Flat	
62% (174)		38% (106)		76% (315)		24% (101)	
consistent		consistent		consistent		consistent	
Yes	No	Yes	No	Yes	No	Yes	No
86% (150)	14% (24)	67% (71)	33% (35)	90% (282)	10% (33)	65% (66)	35% (35)
78% (136)	22% (38)	70% (74)	30% (32)	86% (270)	14% (45)	68% (69)	32% (32)

Notes: N=696; subjects with belief of exactly 60 omitted; upper percentage number for beliefs in the Steep scheme and lower for beliefs in the Flat scheme

Table A5: *Consistent Beliefs - men*

subjects with output<60				subjects with output>60			
41% (152)				59% (217)			
choose Steep		choose Flat		choose Steep		choose Flat	
64% (97)		36% (55)		81% (176)		19% (41)	
consistent		consistent		consistent		consistent	
Yes	No	Yes	No	Yes	No	Yes	No
91% (88)	9% (9)	71% (39)	29% (16)	93% (164)	7% (12)	59% (24)	41% (17)
82% (80)	18% (17)	76% (42)	24% (13)	89% (157)	11% (19)	61% (25)	39% (16)

Notes: N=369; subjects with belief of exactly 60 omitted; upper percentage number for beliefs in Steep and lower for beliefs in Flat

Tables A7 and A8 report standardized dominance statistics for Tables 1 and 2, respectively. Beliefs have by far the highest predictive power in both instances.

Table A6: *Consistent Beliefs - women*

subjects with output<60 39% (128)				subjects with output>60 61% (199)			
choose Steep 60% (77)		choose Flat 40% (51)		choose Steep 70% (139)		choose Flat 30% (60)	
consistent		consistent		consistent		consistent	
Yes	No	Yes	No	Yes	No	Yes	No
81% (62)	19% (15)	63% (32)	37% (19)	85% (118)	15% (21)	70% (42)	30% (18)
73% (56)	27% (21)	63% (32)	37% (19)	92% (113)	8% (26)	73% (44)	27% (16)

Notes: N=327; subjects with belief of exactly 60 omitted; upper percentage number for beliefs in the Steep scheme and lower for beliefs in Flat

Table A7: *Determinants of Gender Sorting - Dominance Analysis of Table 1*

	Standardized Dominance Statistic		
Independent Variable	All	Women	Men
Beliefs	.719	.547	.847
IQ	.129	.183	.065
Extraversion	.046	.032	.055
Emotional Stability	.039	.094	.008
Risk	.028	.101	.002
Female	.020		
Conscientousness	.007	.005	.017
Openness	.005	.032	.002
Agreeableness	.004	.002	.004
High Education	.001	.005	.002

Notes: this table reports on the dominance analysis for Table 1, the determinants of gender sorting; dominance analysis determines the relative importance of independent variables in an estimation model based on contribution to an overall model fit statistic; the general model is the probit regression also used in Table 1, and the model fit statistic is the Pseudo R^2; we see complete general dominance by subject's beliefs

Table A8: *Weighted Determinants of Gender Sorting - Dominance Analysis of Table 2*

	Standardized Dominance Statistic		
Independent Variable	All	Women	Men
Beliefs	.701	.551	.808
IQ	.133	.147	.089
Extraversion	.053	.039	.061
Risk	.035	.117	.003
Emotional Stability	.031	.094	.007
Female	.019		
Openness	.010	.035	.003
Agreeableness	.009	.004	.005
Conscientousness	.009	.006	.024
High Education	.002	.006	.001

Notes: this table reports on the dominance analysis for Table 2, the determinants of gender sorting with weights; dominance analysis determines the relative importance of independent variables in an estimation model based on contribution to an overall model fit statistic; the general model is the probit regression also used in Table 1, and the model fit statistic is the Pseudo R^2; we see complete general dominance by subject's beliefs

Figure A4 reproduces Figure2 and shows gender sorting for inconsistent subjects only. We see that there is no gender sorting effect present that could be explained by inconsistent subjects' beharior: Highly productive women who are inconsistent in our sample even choose Steep *more* often than men (although not significantly so).

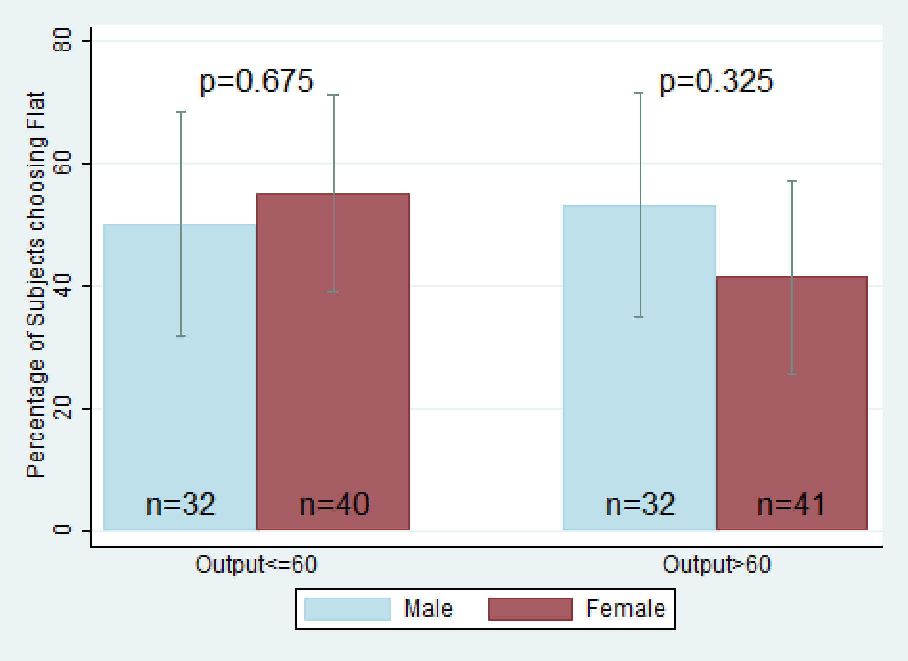


Figure A4 *This figure replicates Figure 2 (sorting into payment schemes) for inconsistent subjects only; no significant differences; p-values for a WRST; N=145*

A.4 Robustness Checks

In Section 3.1 we saw that highly productive women more often sort into Flat than men, and we have seen in section 3.2.A that beliefs map into choice. It follows that women with high outputs should have lower beliefs than their male counterparts. This is confirmed in Figure A5, where we expand upon Figure 4 and add output to the graph. The percentage numbers of men/women in a section are shown in the corners. We see that for many subjects, productivity beliefs are far off their actual output, as the 45° line signifies matching beliefs and output; any subjects above (below) the line are overconfident (underconfident). Moreover, men have more variance in their distribution and there are more men than women who have very low outputs.

Subjects in the upper right and lower left parts of the figure correctly anticipate their output and therefore sort efficiently. Beliefs for subjects in the upper left and lower right lead them to wrong decisions about their payment scheme, as they overestimate or underestimate their productivity. Due to their incorrect anticipation of own productivity, subjects in the lower right wrongly sort into Flat and therefore make a type II sorting error. From the percentages we can see that highly productive men are much less likely to have false low beliefs than women ($p < .001$, WRST), as only 5.6% of highly productive men have low beliefs compared to 14.0% of women. This result then explains why women too often incorrectly sort into Flat, as highly productive women more often have low beliefs about their productivity than men. Figure A6 is the same as Figure A5, but inconsistent subjects are included; we see no qualitative differences to the distribution in Figure A5. The gender difference is still significant in quadrant IV ($p < .001$, WRST) and nowhere else, and we find no gender differences in consistency within quadrants (all $p > .138$, WRSTs).

Table A9 is the same as Table 5, income of our subjects in the labor market, but includes inconsistent subjects. Results stay qualitatively the same, as beliefs are still a highly significant predictor for highly productive women's income.

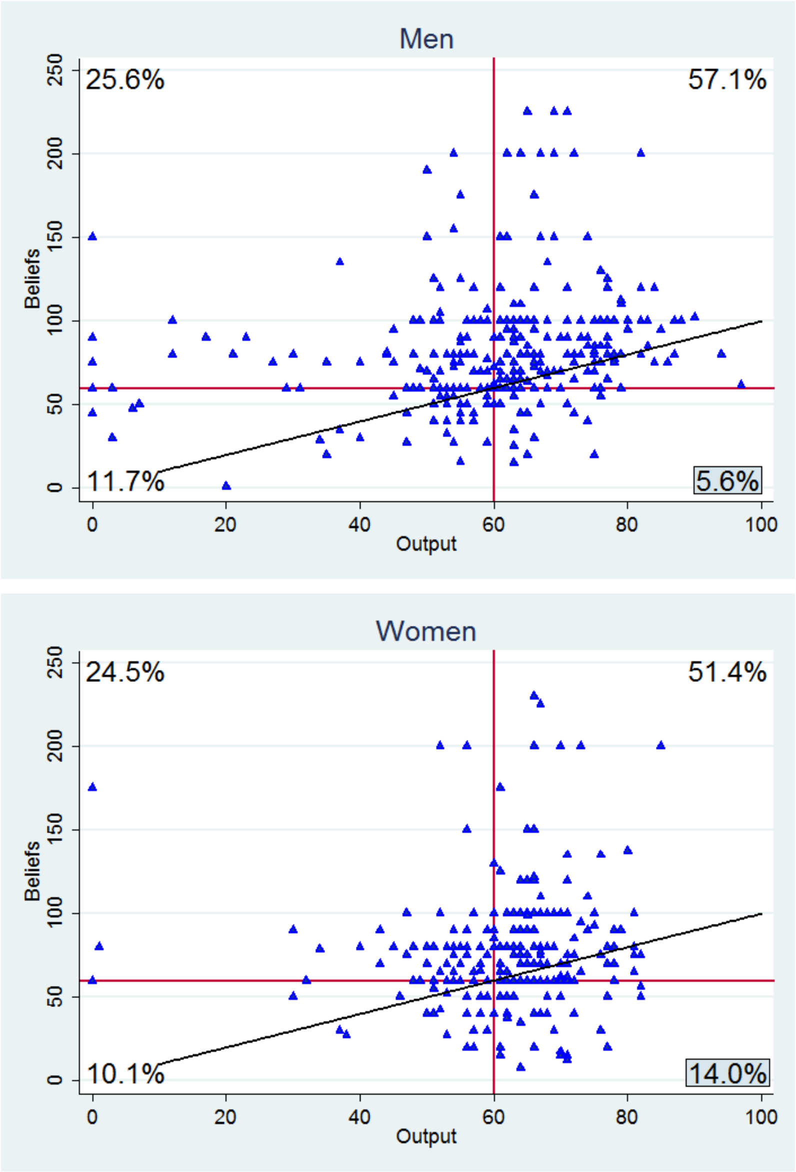


Figure A5 *Scatterplot of output and beliefs for men and women separately; significantly more women in lower right quadrant IV; numbers in corners indicate the percentage of men/women in that quadrant, omitting anyone with beliefs or productivity of exactly* 60 *and inconsistent subjects; lines at the break-even points of 60; 45° line indicates matching ex-ante beliefs and ex-post output; N=570 (314 men and 256 women)*

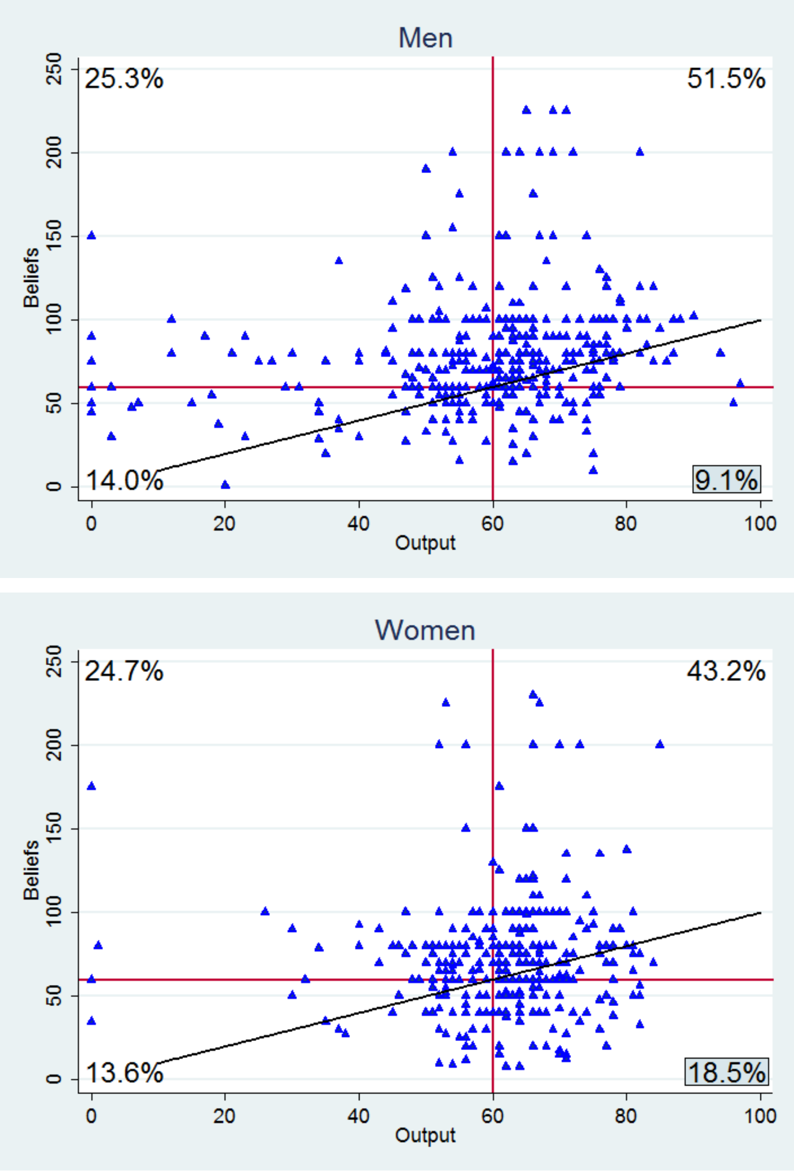


Figure A6 *Robustness check for Figure A5; inconsistent subjects included; significantly more women in lower right quadrant IV; numbers in corners indicate the percentage of men/women in that quadrant, omitting anyone with beliefs or productivity of exactly* 60*; lines at the break-even points of 60; 45° line indicates matching ex-ante beliefs and ex-post output; N=715*

Table A9: *Income - Robustness*

	Output< 60		Output> 60	
	Men	Women	Men	Women
Beliefs < 60	$-.449^{*}$	$.152$	$-.092$	$-.587^{***}$
	$(.240)$	$(.203)$	$(.199)$	$(.161)$
Output	$-.011$	$.003$	$.004$	$.018$
	$(.007)$	$(.008)$	$(.010)$	$(.013)$
IQ	$-.028$	$-.072$	$.026$	-0.010
	$(.038)$	$(.036)$	$(.028)$	$(.026)$
High Education	$.439^{*}$	$.388^{*}$	$.249$	$.105$
	$(.246)$	$(.232)$	$(.159)$	$(.169)$
Age	$.069^{*}$	$.150^{***}$	$.187^{***}$	$.120^{***}$
	$(.040)$	$(.037)$	$(.030)$	$(.032)$
Age$^2/100$	$-.048$	$-.143^{***}$	$-.169^{***}$	$-.106^{***}$
	$(.039)$	$(.037)$	$(.032)$	$(.035)$
Risk	$-.024$	$-.038$	$.002$	$.032$
	$(.044)$	$(.031)$	$(.026)$	$(.026)$
Constant	1.466	$-.666$	-2.413^{**}	-1.293
	(1.095)	(1.014)	$(.937)$	(1.221)
Adj. R^2	$.110$	$.233$	$.235$	$.208$
NObs	124	106	193	173

*Notes: Robustness check for Table 5; inconsistent subjects included; OLS regressions; dependent variable is the log of annual gross income from tax reports in DDK divided by 10,000; highly productive women of high beliefs have higher incomes; "Beliefs < 60" is a dummy variable; "High Education" takes the value 1 if subject has a Bachelor degree or higher; "Risk" means the number of safe choices in a Holt/Laury test; standard deviation below the coefficients in parentheses; ***:* $p < .01$, ***:* $p < .05$, **:* $p < .1$

A.5 Instructions

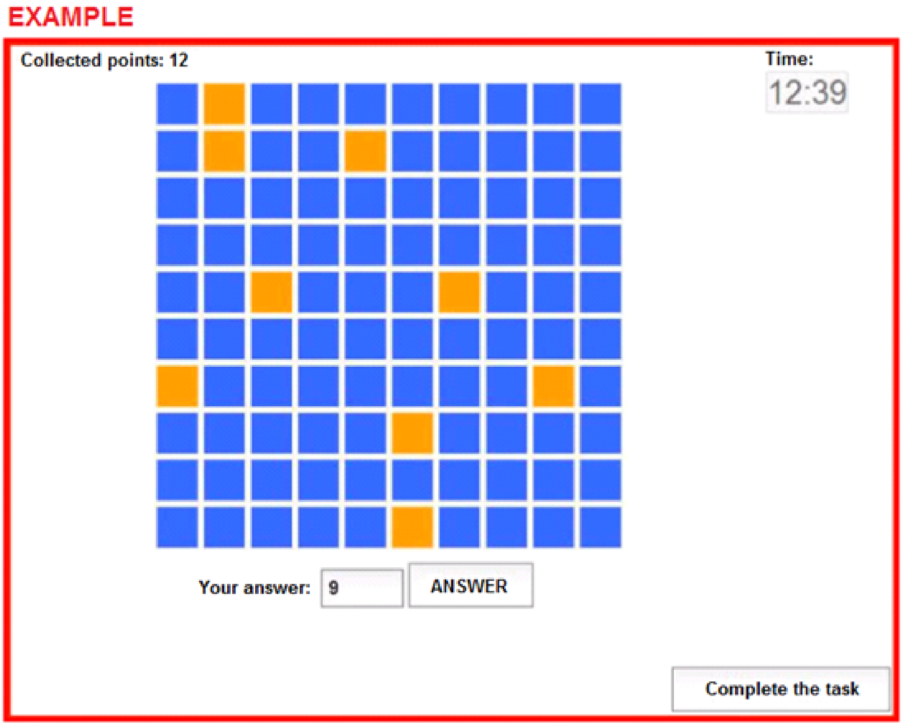
We reprint translated instructions below. Original instructions were in Danish.[27]

Screen 1:

In this part of the experiment you have 15 minutes to work on a task to earn money. The next screen will show you what the task is about.

Screen 2:

Here you see an example of the task:



EXAMPLE

Collected points: 12

Time: 12:39

Your answer: 9

Figure A7 *Task example*

You will see 10-by-10 grids which contain blue and yellow squares, just like in the example above. The amount of yellow squares varies from one figure to the next. Your task is to count the number of yellow squares in the figure. You earn 1 point per correct answer. When your answer is correct, which is 9 in the example, the next grid will be shown.

[27] Taken from subject 123828, screens 322pp in iLEE3.

You have 15 minutes to work on the task. A countdown timer in the upper right corner of the screen shows how much time is left. Be aware of the fact that once you start working on the task, the countdown timer is independent of whether you take a break. If you close your browser and log in after the 15 minutes have passed, you will not continue from the point where you left, but you will be taken to the next screen in the experiment.

This task becomes increasingly complicated the more right answers you provide. If you wish to finish the task before the time has run out, then press the End task button in the lower right corner of the screen. When you end the task, or when the time runs out, this part of the experiment is over. The task does not start before a few screens. The next screen will explain how your income for the task is determined.

Screen 3:
This is how your income is determined
Your earnings from this part of the experiment depend on the number of points that you collect.

You will work under one of these income systems:

Income System 1:
You get 1,5 kr. for each point you collect.
Income System 2:
You get 60 kr. irrespective of the points you collect, plus 0,5 kr. for each point you collect.

Note that under both payment systems, the more points you collect, the more you will earn. If you collect more than 60 points, you will earn more under income system 1. If you collect less than 60 points, you will earn more under income system 2.

Example 1:
Assume that you collect 80 points.
Under income system 1, you get : 1,5 kr. * 80 points = 120 kr.
Under income system 2, you get : 60 kr. + 0,5 kr. * 80 points = 100 kr.

Example 2:
Assume that you collect 40 points.
Under income system 1, you get : 1,5 kr. * 40 points = 60 kr.
Under income system 2, you get : 60 kr. + 0,5 kr. * 40 points = 80 kr.

Screen 4:
How many points do you think you will collect compared to others?

Before you will be informed about the income system that you will be working under, you are asked to indicate your beliefs about your own productivity in comparison to the other participants in the experiment under the two different income systems. The beliefs you state here have no influence on the income system under which you will work later.

Income System 1: You get 1,5 kr. for each point you collect. If all participants work under income system 1, do you think you will be among the fifth of the participants, who are:[28]

The least productive - at least 80% will collect more points than you.
The second least productive - at least 60% will collect more points than you, and at least 20% will collect less points than you.
Averagely productive - at least 40% will collect more points than you, and at least 40% will collect less points than you.
The second most productive - at least 20% will collect more points than you, and at least 60% will collect less points than you.
The most productive - at least 80% will collect less points than you.

Income System 2:

You get 60 kr. irrespective of the points you collect, plus 0,5 kr. for each point you collect. If all participants work under income system 2, do you think you will be among the fifth of the participants, who are:[29]

[28] Subjects had to click on radio buttons to indicate their expectation.
[29] Subjects had to click on radio buttons to indicate their expectation.

The least productive - at least 80% will collect more points than you.
The second least productive - at least 60% will collect more points than you, and at least 20% will collect less points than you.
Averagely productive - at least 40% will collect more points than you, and at least 40% will collect less points than you.
The second most productive - at least 20% will collect more points than you, and at least 60% will collect less points than you.
The most productive - at least 80% will collect less points than you.

Screen 5:
How many points do you think you will collect?

You are being asked to indicate the amount of points that you believe you will collect under each income system. The beliefs you state here have no influence on the income system under which you will work later.

I believe I will collect ________ points under income system 1 (1,5 kr. per point).
I believe I will collect ________ points under income system 2 (60 kr. for sure, plus 0,5kr. per point).

Screen 4:
How the income system is chosen
All participants in the experiment can choose the income system they prefer. Half of the participants will be working in their preferred income system. For the rest of the participants it is determined at random which income system they work under. Both systems are equally likely. This means that you have a 75% chance to work under your preferred income system.
You will be informed about your income system on the next screen.

Which income system do you choose?[30]

Income System 1: You get 1,5 kr. for each point you collect.
Income System 2: You get 60 kr. irrespective of the points you collect, plus 0,5 kr. for each

[30] Radio buttons were available to indicate the choice.

point you collect.

Screen 7:
The income system has now been decided. You are among the participants who work under their preferred income system.
You chose income system 1. You will get 1,5 kr. for each point you collect.
It is now time to accumulate points. You can re-read the instructions for the task by pressing Revisit Instructions in the upper right corner of the screen.
When you are ready to start on the work task, press Start. You will then have 15 minutes to collect points.

Screen 8:
Work Task[31]

Screen 9:
This experiment is now over
You collected xx points. You get 1,5 kr. per point. Your earnings from points collection is therefore yy DKK.

[31] Like in screen 2.

Distribution of Cells

Figure A8 shows the distribution of yellow cells to count over time. The number of yellow cells goes up with the total number of competed tasks in a noisy fashion, and reaches about 45 after about 80 completed tasks. Participants were made aware that the task gets harder over time, but were not informed about the distribution itself.

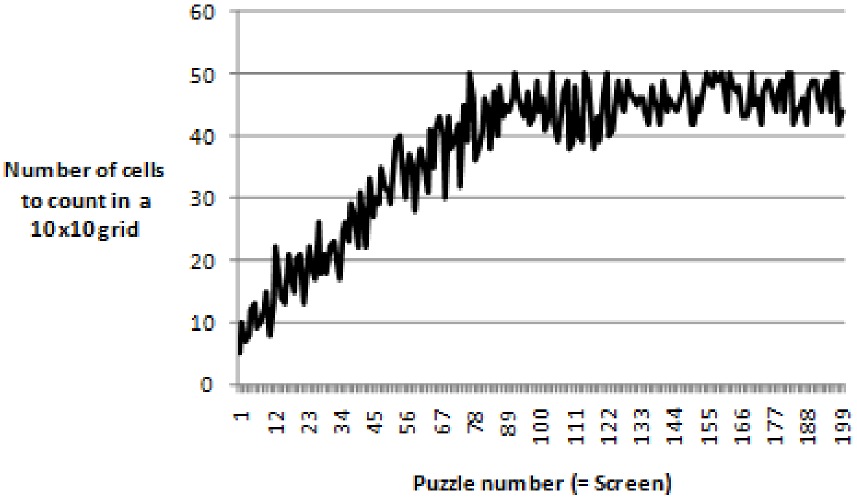


Figure A8 *Distribution of yellow cells to count in the work task*

Abstracts

My thesis consists of three papers, which are all experimental in nature. In the first paper we give a holistic overview and comparison of all risk elicitation methods used today, and recommend that a derivation of Holt and Laury's (2002) well-known method, where the high payoff instead of the probabilites are varied, should be used as it emerges as the best method in the dimensions of forecast accuracy and stability. In the second paper I analyze a three-player variant of Hotellings (1929) influential locational choice model experimentally and theoretically, where entry is sequential instead of simultaneous. I find that first, Duverger's Law (1959) is robust to violations of the gametheoretical prediction also in this context, and second, that special attention should be directed towards the timeframe of experiments, as results vary massively with different timeframes for our subjects. While initial play is not according to theory at all, after many rounds play converges towards the theoretical prediction. The third paper investigates the role of beliefs on job choice in the labor market, and we find in an experiment that women of high productivity too often sort into jobs that would be ideal for workers of low productivity. We identify wrong productivity beliefs as the source of this detrimental sorting, and we also find that this effect translates to the Danish labor market, where women of high ability but low confidence earn less than their equally productive peers.

Meine Dissertation besteht aus drei experimentellen Papers. Im ersten Paper geben wir einen Überblick und einen Vergleich über alle Risikomessmethoden, welche heute gebräuchlich sind. Als Ergebnis empfehlen wir, eine Methode zu wählen, welche eine Variante der in der Literatur gebräuchlichsten Methode von Holt und Laury (2002) ist, mit dem Unterschied, dass die höchsten Auszahlungen und nicht die Wahrscheinlichkeiten verändert werden. Diese Variante stellt sich als die beste in den beiden wichtigsten Dimensionen Vorhersagekraft und Stabilität heraus. Im zweiten Paper analysiere ich eine Variante des einflussreichen Modells von Hotelling (1929) für drei Spieler mit sequentiellem Eintritt. Dabei erweist sich einerseits Duvergers Gesetz gegenüber Verletzungen der spieltheoretischen Vorhersagen auch in diesem Kontext als robust. Andererseits spielt es eine große Rolle, wie lange man Subjekte im Experiment lernen lässt, da sich die Resultate mit unterschiedlichem Zeitrahmen massiv ändern. Obwohl am Anfang des Spiels das Verhalten der Subjekte nicht mit der theoretischen Vorhersage übereinstimmt, konvergiert es nach vielen Runden zu dem theoretisch vorhergesagten. Das dritte Paper untersucht experimentell die Rolle von Selbsteinschätzung auf die Wahl eines Jobs. Wir stellen

fest, dass hochproduktive Frauen zu häufig Jobs wählen, die für niedrigproduktive Arbeiter ideal wären. Als Grund isolieren wir falsche Vorstellungen über die eigene Produktivität. Der Effekt, dass hochproduktive Frauen mit geringem Vertrauen in die eigenen Fähigkeiten falsche Jobs bevorzugen, überträgt sich auch auf den dänischen Arbeitsmarkt. Auch hier haben hochproduktive Frauen, die ihre Fähigkeiten unterschätzen, niedrigere Einkommen gegenüber jenen, die ihre Fähigkeiten korrekt einschätzen.