Essays on Forecasting Curves and Behavioral Economics

Dissertation zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft

eingereicht an der Fakultät für Wirtschaftswissenschaft der Universität Regensburg

vorgelegt von: Alexander Lauf

Berichterstatter:

Prof. Dr. Andreas Roider

Prof. Dr. Rolf Tschernig

Tag der Disputation: 15. November 2022

Universität Regensburg

Dissertation

Essays on Forecasting Curves and Behavioral Economics

Alexander Lauf

Inauguraldissertation zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften eingereicht an
der Fakultät für Wirtschaftswissenschaften der Universität Regensburg

Erstbetreuer: Prof. Dr. Andreas Roider
Zweitbetreuer: Prof. Dr. Rolf Tschernig

März 2022

Für Anita

Acknowledgements

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

I would first like to extend my deepest gratitude to my first supervisor Professor Andreas Roider for his guidance and support during the work on my dissertation. You always had an open ear for my research ideas and organizational matters. You made it possible for me to pursue my research interests. I have benefited substantially from your comments on experimental designs and paper drafts.

I'm also extremely grateful to my second supervisor Professor Rolf Tschernig for his continued support and encouragement. You have always taken an extraordinary amount of time to discuss my projects and provided very valuable comments. I greatly benefited from your experience and technical accuracy in all methodological matters.

I further want to thank my colleagues from the University of Regensburg. In internal presentations and discussions about projects and experimental designs, I received great feedback and comments, which helped me to improve my work. Also, the atmosphere within this group was always very collaborative and cheerful.

I would particularly like to thank the colleagues I jointly worked on projects - Lars for our long insightful discussions about everything, Silvio for the great project initiative, Vanessa for all the joint data work, and Helena for her manly spirit driving the project.

At this point, I also want to emphasize how much I profited from my membership in the International Graduate Program "Evidence-Based Economics" (EBE) of the Elite Network of Bavaria. The EBE syllabus was very well structured and contributed significantly to my success. Through the EBE, I had the opportunity to exchange ideas with many interesting people. Additionally, I would like to acknowledge the generous funding which made it possible for me to conduct the experiments in my thesis.

Moreover, I would like to thank my family for their support and sympathetic ear. You are always there for me. Finally, I could not have completed this dissertation without the support and patience of my fiancée, Anita, who unconditionally supported me throughout the whole time. Thank you!

Alexander Lauf, 2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *By definition all scientists are data scientists. In my opinion, they are half hacker, half analyst, they use data to build products and find insights. It's Columbus meet Columbo – starry-eyed explorers and skeptical detectives.*
>
> MONICA ROGATI

In a world where reliable information is becoming increasingly important but also needs to be available more and more quickly, flexible and accurate forecasting techniques are indisputably relevant. Forecasting is the procedure of making predictions about future tendencies and events using past records. The approaches range from more or less elaborated guesses to advanced statistical methods all the way up to state-of-the-art techniques using massive data sets, artificial intelligence, and vast amounts of resources. Even though predictions are an excellent basis for qualified decision-making, the analysis of causal relationships should not be neglected either, as it is a central component for the conception and elaboration of the next course of action. Causal analysis requires a controlled environment, and there are no better methods than conducting experiments. The increasing digitalization has also opened up new possibilities for the implementation of experiments. Over time, a variety of platforms have been established that are successfully used by economists. Some well-known names are Amazon MTurk and Prolific, which were also used in this thesis. This dissertation consists mostly of three projects that can be assigned to two parts, each with a shared general topic. The first part consists of Chapter 2, which presents a glimpse at selected methods, and Chapter 3, which outlines the results of the first project in the field of forecasting.

The first project aimed to find a new approach to predict supply curves based on historical data by a highly specialized neural network. Such supply curves typically emerge from auctions in the German balancing market. The balancing market is an essential part of the energy market and the energy security's cornerstone. The underlying idea of the presented method is to model curves as sequences and utilize an optimized recurrent neural network for sequence-to-sequence prediction while imposing an autoregressive structure on the input data. In an application to the German balancing power market, the proposed method provides more accurate forecasts than classical or functional data forecasting approaches. The presented approach also allows addressing the problems of varying sequence lengths and irregularly spaced observation periods. It can easily be extended to other fields or augmented for the inclusion of covariates. Besides that, it provides the appealing feature of visualizing the operating mode of a machine learning method in the form of an attention plot.

The second part of the dissertation presents the results of the economic experiments conducted in the second and the third project. The most significant advantage of using experiments to measure economic and social behavior (rather than purely observational data) is that the direct impact of one or more factors can be identified because of the randomization and controlled environment. Conducting economic experiments is also subject to specific international standards. These include giving the participant monetary incentives and avoiding any form of deception. Also, all manuals, questionnaires, and protocols of the experimental procedure are published upon publication of the project and are generally accessible. This ensures that the experimental method is transparent and allows for replication and verification of the studies by other research groups. The experimental method is recognized as the “gold standard” in collecting data on social and economic behavior, as also evinced by the award of the Nobel Prize to economists for their research with economic experiments.

Chapter 4 presents the results of the second project. The project’s goal was to study the role of round numbers in bargaining situations. Recent years have seen a growing body of literature on the effect of round numbers in decision-making. This study focuses on whether this effect can be explained by preferences for round numbers (*round-number bias*) or by their role as a solution for a coordination problem (*focal point*). To this end, it is analyzed how these two channels relate to round-number clusters in observational and experimental data on price negotiations. It was hypothesized that faster decisions and higher acceptance frequencies result from a round-number bias, focal points, or both. In a first step, using data from Backus et al. (2020), it is found that a large fraction of successful negotiations end with round prices and that round prices correlate with faster agreements. In a second step, an experiment to disentangle the channels is designed. The study was conducted on Amazon MTurk and confirms that round numbers are associated with quicker decisions. Moreover, evidence for the relevance of both channels - bias and focal points is found.

Chapter 6 presents the results of the third project. The object of the project was to understand how much of the differences in the behavior of men and women often found in the economic literature can really be associated with gender as opposed to an individual’s sex. This question is investigated by using well–known behavioral economic experiments in the domain of competitiveness, risky choices, and altruism. Gender has come out to be a key factor explaining differences in behavior. This project follows a systematic approach to test for gender and sex differences in behavior. An experiment is conducted where first correlations of gender and sex with competitiveness, risk–taking, and altruism are analyzed by comparing decisions of cisgender (cismen and ciswomen) and transgender (transmen and transwomen) individuals. Second, the participants are primed with either a masculine or a feminine gender identity. By subconsciously activating a gender, a causality between gender and behavior in our sample of cis- and transgender participants can be established. It is hypothesized that if gender (and not sex) is indeed a primary factor for decision–making, (i) individuals of the same gender (and different sex, i.e., ciswomen/transwomen and cismen/transmen) make similar decisions, and decisions significantly differ when gender differs (and sex is the same, i.e., cismen/transwomen and ciswomen/transmen), and (ii) priming changes behavior.

The dissertation is structured as follows. Chapter 2 covers selected aspects of the methodology. Chapter 3 presents the first project. Chapter 4 provides the results of the second project. Chapter 5 gives an introduction to the literature on gender differences. Chapter 6 describes the third project. Finally, Chapter 7 concludes.

Chapter 2

Methodology

We live in a world of big data that empowers businesses, organizations, and people to make data-driven decisions – intentionally or unintentionally. Modern devices record activities twenty-four hours, seven days a week, regardless of the activities' analogous or digital nature. Smartphones track our hiking routes or collect our search history on the web. Smart home furniture listens to us and abides by our spoken commands. The daily life examples highlight the vast number of data sources and the need for methods to handle big data. The data are commonly stored as time series. Of course, time series are not only found in everyday situations but also in the economic world, e.g., as market indicators, sales numbers, or stock prices.

This chapter introduces a selection of methods that were used in the subsequent chapters. They allow handling particular forms of data. Section 2.1 is dedicated to bringing discrete observations into a smoothed form and provides a short application. Section 2.2 briefly discusses test procedures when the data is not recorded as expected. The informed reader is free to skip these sections and start directly with chapter Chapter 3.

2.1 From discrete observations to smooth curves

As real-world observations almost always include some form of obstacles, researchers have to make sure that the data meets all requirements for the methods they want to apply. For example, suppose an investor has been watching some stocks on the market for several years. However, since the enterprising investor does not have that much time, he only looks at the current price once every day. The prices for one stock over time form a time series. The collection of all these time series and, therefore all stocks could be a functional data set. Ramsay (1982) and Ramsay and Dalzell (1991) coined the term *functional data analysis* (FDA) that operates on such data sets. It assumes that the data are observed in functional form. However, since information is commonly collected discretely over time (e.g., daily as in the stock price example) and probably also includes noise, the first step of FDA is smoothing, turning the observed data into smooth curves.

2.1.1 Basis functions

One popular method is *B-Spline smoothing* which requires an introduction to basis function systems and spline functions. An excellent reference is De Boor (2001). In this context, a smooth function is a function that possesses one or more derivatives. The goal is to find a smooth function g for $y = g(x)$ in the case that only n discrete data points of the form (x_i, y_i) for $i = 1, ..., n$ are observed.

Ramsay and Silverman (2005) define a basis function system as "*a set of known functions that are mathematically independent of each other and that have the property [...] (to) approximate arbitrarily well any function by taking a weighted sum or linear combination of a sufficiently large number J of these functions*"(p.43). Hence, with this property, a linear combination of J basis functions can represent the function looked for. If the number of basis functions equals the number of observations ($J = n$) it is called *interpolation*. In the other case ($J < n$), it is named *smoothing*. So, the smoothing depends on the number of basis functions and the choice of their form.

There are multiple forms of basis functions. Besides the Fourier, polynomial, exponential, and power bases, the *spline basis* is a widespread choice. The term spline refers to a long and flexible strip of wood, plastic, or metal that draftsmen used to draw a smooth curve by fixing the strip at specific points and bending it in between. The aim is to approximate $g(x)$ over the interval $[\tau_0, \tau_L]$. Additionally, there are $L - 1$ knots τ_l with $l = 1, \ldots, L - 1$ that separate the interval $[\tau_0, \tau_L]$ in L subintervals. Over each interval, the m-order spline is a polynomial of degree $m - 1$. The degree refers to the polynomial's highest power. Let $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_{L-1}, \tau_L)$ be the non-decreasing knot sequence, where τ_0 and τ_L are referred to as boundary knots or endpoints. So, a spline is defined by its order m and the knot sequence $\boldsymbol{\tau}$.

It should be noted that multiples, sums, and differences of a spline function remain a spline function. So, a basis function system of spline functions effectively remains a spline function. Splines are restricted twofold. First, for $m > 1$, adjacent polynomials have to join up smoothly at breakpoints, so that their function values are equal. Second, derivatives up to order $m - 1$ have to be equal at these breakpoints.

De Boor (2001) denotes for a given knot sequence $\boldsymbol{\tau}$ and the knot τ_j the corresponding j-th B-spline of order k by $B_{j,k}$. For $k = 1$, the first-order B-Spline is given by

$$B_{j,1}(x) = \begin{cases} 1, & \text{if} \quad \tau_j \leq x < \tau_{j+1}, \\ 0, & \text{else.} \end{cases} \tag{2.1}$$

These $B_{j,1}(x)$ are constrained to form a partition of unity, i.e., $\sum_j B_{j,1}(x) = 1$. In other words, if two adjacent knots are identical, the B-Spline consists of zeros,

$$\tau_j = \tau_{j+1} \quad \rightarrow \quad B_{j,1}(x) = 0. \tag{2.2}$$

For k-order B-Splines with $k > 1$, the following recurrence relation can be used,

$$B_{j,k}(x) = w_{j,k}B_{j,k-1}(x) + (1 - w_{j+1,k})B_{j+1,k-1}(x), \tag{2.3}$$

with

$$w_{j,k}(x) := \frac{x - \tau_j}{\tau_{j+k-1} - \tau_j}. \tag{2.4}$$

Let K be the highest order of the B-splines, then there are $J = K + L - 1$ basis functions, and the spline function based on B-splines to approximate the target function is given by

$$g(x) = \sum_{j=1}^{J} c_j B_{j,K}(x), \tag{2.5}$$

where c_j are coefficients corresponding to each B-spline.

An illustration is in order. For the online streaming platform `twitch.tv`, Table 2.1 reports the average, rounded numbers of viewers on each weekday for 143 streamers for a two-year period. A streamer is an individual who shares his/hers computer screen with an online audience and engages in interactive chats. Details on this unique data set can be found in Box 1.

Table 2.1. Example: Average number of viewers per weekday.

Weekday	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Index (x)	1	2	3	4	5	6	7
Number of viewers (y)	1519	1760	1814	1874	1573	1436	1322

Note: The number of viewers is the average number for 143 streamers who provided content for one selected topic on twitch.tv from 2016 to 2018.

Since the construction of B-splines depends on the knot sequence, in the very first step, we need to take a closer look at the form of $\boldsymbol{\tau}$. Looking at Table 2.1, it almost occurs naturally that $\boldsymbol{\tau}$ should include all observed values of x. As briefly discussed later for p-Splines, in Section 2.1.3, this approach yields useful properties, and on the practical side, it is easy to implement and circumvents the discussion over a reasonable knot placement. In principle, the knots can be placed as required for the application, e.g., multiple knots at higher curvature, equally or unequally spaced knots, or many and few knots. Therefore, in this exemplary application, the boundary knots are set to the additional indices 0.5 and 7.5. This shifts the center of the day to the integer indices due to the assumption that the curve is continuous in time.

Box 1: `twitch.tv`

twitch.tv is the leading platform for streaming services and was launched on June 6, 2011. Amazon bought it for $970 million in 2014, and its headquarters is based in San Francisco, CA. The website `Alexa.com` ranks twitch.tv 20th in Germany and 37th globally on 2021/09/24 based on a combination of average daily visitors and pageviews over the past month. According to twitch.tv, more than 7 million unique creators stream each month, and the website is visited by more than 30 million visitors daily on average. Moreover, just for 2020, the platform reports that viewers watched one trillion minutes.

On twitch.tv, streamers share their computer screen with an audience of viewers via the internet. Registering is free of charge, and the audience can interact with the streamer by chat while the streamer commonly shares webcam footage and uses a microphone to talk with the viewers. In addition, viewers can opt to follow streamers to receive constant updates or purchase a paid subscription to support them financially.

The primary measure of success in the entertainment industry is the audience rating. The number of viewers crucially determines how attractive a show is and how many people are willing to spend their time and attention on it. Creators compete for viewers by flashy channel labels, creative hints about the stream's quality, or by incentivizing with offering giveaways, besides many more methods. Hence, it is also an attractive field for research in economic behavior.

The website `SullyGnom.com` created and maintained by David records activities on twitch.tv and has provided aggregated data since 2015. With David's help, I collected a unique data set for the period of 2016/05/24 to 2018/06/30 for 143 streamers and their audience size, amounting to $n = 543,850$ observations with a resolution of 15 min if a stream was active.

To formally summarize, the example consists of $n = 7$ observations. We approximate the function on the interval $[0.5, 7.5]$, which covers the index, x, from 1 to 7. At each data point and the boundaries a knot is placed, leading to the knot sequence $\tau = (0.5, 1, 2, 3, 4, 5, 6, 7, 7.5)$. So, there are $L = 8$ subintervals between $\tau_0 = 0.5$ and $\tau_L = 7.5$. We use cubic B-splines, i.e., the order is $K = 4$, and the degree is $K - 1 = 3$, so piecewise cubic polynomials cover the intervals. The number of basis functions is $J = 4 + 8 - 1 = 11$.

The procedure to obtain the basis function consists of three steps. In the first step, the knot sequence is extended by adding the boundary knots as many times as the degree of the B-spline. So, the knot τ_0 is prepended to the knot sequence, and the last knot, τ_L, is appended $K - 1$ times. In the example, this leads to,

$$\tau = (0.5, 0.5, 0.5, \quad 0.5, 1, 2, 3, 4, 5, 6, 7, 7.5, \quad 7.5, 7.5, 7.5).$$

This extension accounts for the fact that the function may be discontinuous beyond the boundaries. The restrictions on the splines normalize the derivatives to 0, and taking a closer look at the denominator of Eq. (2.4) the last $K - 1$ repeated boundary knots guarantee the computability of the weight, $\omega_{j,k}$.

The following steps go along the sequence $k = 1, ..., K$ and form an iterative process. In the second step, the first-order B-splines ($k = 1$) defined in Eq. (2.1) are computed.

These B-splines form step functions equaling 1 over the intervals between two knots and 0 otherwise. A visualization of the basis functions helps to follow.[1]

In Fig. 2.1, the top panel shows the step function, $B_{2,1}$ for the time between Monday and Tuesday in Table 2.1. The B-splines for the remaining seven intervals between the dotted lines are constructed the same way.

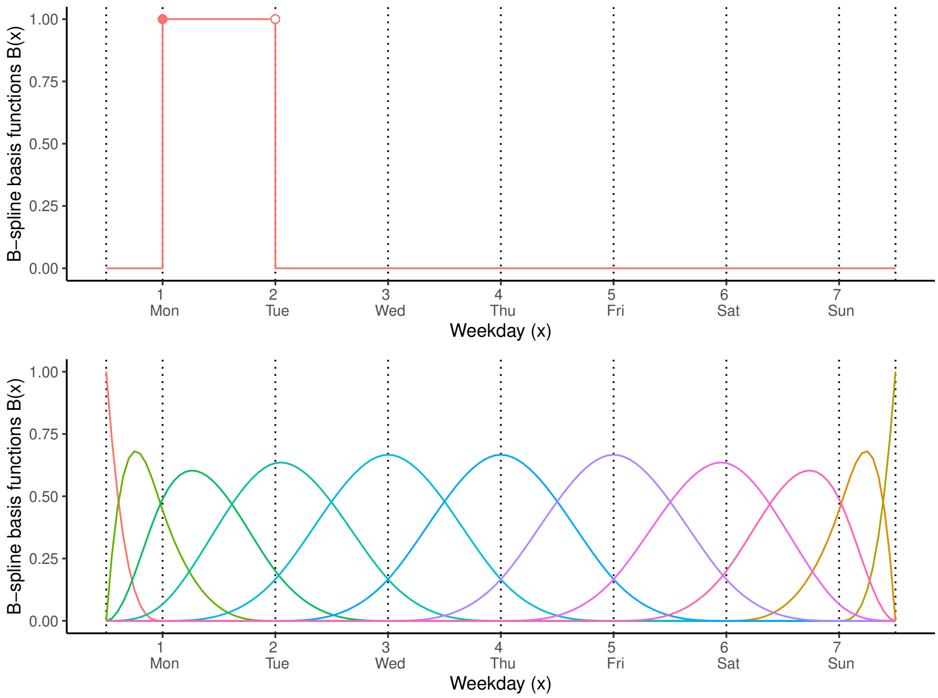


Figure 2.1. Basis functions.

Note: The top panel shows the 2nd of the $J = 8$ first-order B-spline ($k = 1$) basis functions, $B_{j,1}(x)$, defined in Eq. (2.1). The dot marks the included point while the circle the excluded point. The bottom panel shows the $J = 11$ fourth-order or cubic B-spline ($k = 4$) basis functions, $B_{j,4}(x)$. The knot sequence, $\boldsymbol{\tau}$, is identical for both panels. The dashed lines mark the knots. The coloring highlights the j-th B-spline basis function. All basis functions are evaluated on points equally-spaced by 0.05 over $[\tau_0 = 0.5, 7.5 = \tau_L]$.

In the last and third step, the recurrence relation described in Eq. (2.3) for $k > 1$ is iteratively applied to the previously constructed step functions. The application of Eq. (2.3) is repeated $K - 1$ times and allows to obtain the $J = 11$ cubic B-spline basis functions, which are shown in the bottom panel of Fig. 2.1.

The bottom panel of Fig. 2.1 allows to highlight a few characteristics. First, the vigilant reader might have noticed that the construction of the basis functions only depends on x and the knots, and so far, y is not included. This will be the topic covered in the next section. Secondly, the basis functions are non-negative over the whole interval between

[1] There are abundant possibilities to work with splines as they are implemented in any major mathematical or statistical software tool. For example, the advances and application in medical research and biostatistics led to a recent overview by Perperoglou et al. (2019) for `R`. However, I provide a short routine to illustrate the procedures since the actual calculations are deeply nested within sub-routines or other packages for efficient programming and computing reasons. It is online reachable by `Link`, or Online (2022k).

the boundary knots. Moreover, each basis function is non-zero for at most K adjacent intervals. So, it has compact support. Thirdly, for each point, the function values of the basis functions add up to 1. This follows from the construction of the basis function and the restriction to form a partition of unity. Fourthly, drawing a vertical axis of symmetry on Thursday ($x = 4$) shows that the left and right sides are identical. The shape of the central basis functions with their peaks on Wed, Thu, and Fri are identical. The others are similar but differ in their transition to zero on the left or right side, respectively. It is the result of the repeated knot placement at the end of the knot sequence. Overall, all splines have a smooth transition to zero since cubic B-splines have two continuous derivatives.

2.1.2 Smoothing spline

In order to obtain an estimate for the function g, we now need to incorporate the observed y. For this purpose, a powerful method is the *smoothing spline* that introduces a roughness penalty and a smoothness parameter λ that governs the smoothness of the fit. The trade-off between a fit term and a smoothing term can be summarized in the *penalized residual sum of squares* that is

$$\mathcal{L}_\lambda = \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(x)^2 dx, \tag{2.6}$$

where g is given by Eq. (2.5), and g'' denotes the second derivative with respect to x. For a given λ, the spline estimate $\hat{g}_\lambda$ of g is the function that minimizes $\mathcal{L}_\lambda$. Using cubic smoothing splines guarantees that g'' exists. In Eq. (2.6), the sum over the discrete sampling points represents the fit term of how close the estimate matches the data, while the integral measures its roughness and acts as penalty on $\mathcal{L}_\lambda$. A cubic spline with knots at each data point x_i minimizes $\mathcal{L}_\lambda$, as proven in Chapter V of De Boor (2001). The smoothing parameter, λ, can be found by the *generalized cross-validation* (GCV) criterion proposed by Craven and Wahba (1978),

$$\mathrm{GCV}(\lambda) = \frac{n \sum_{i=1}^{n} (y_i - \hat{g}_\lambda(x_i))^2}{(n - df(\lambda))^2}, \tag{2.7}$$

where $df(\lambda)$ denotes the degree of freedom which is the trace of the sub-projection operator $\mathbf{S}_\lambda$ that satisfies, $\hat{\mathbf{y}} = \mathbf{S}_\lambda \mathbf{y}$, where $\mathbf{y}$ denotes the vector of discrete data to be smoothed. So, $df(\lambda) = \operatorname{trace} \mathbf{S}_\lambda$ (for details, see Ramsay and Silverman (2005)). The search for the optimal λ requires trying multiple values for λ. With $\lambda = 0$, the *regression spline* is obtained, which is a non-penalized smoothing spline.

2.1.3 P-spline

The estimation of Eq. (2.6) requires approximating the integrated squared derivative. O'Sullivan (1986) proposed to use a penalty based on the coefficients and second derivative of the fitted curve. Eilers and Marx (1996) extended this idea and introduced the *penalized splines* (P-splines). P-Splines do not depend on integrals or derivatives but use a purely discrete smoothing term that utilizes q-th order difference operators on the coefficients. Hence, their computation can be easily handled. Formally, the estimate $\hat{g}$ minimizes,

$$\mathcal{L}_\lambda^P = \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \sum_{j=q+1}^{J} (\Delta^q c_j)^2, \tag{2.8}$$

where the difference operator Δ^q is defined so that $\Delta^1 c_j = c_j - c_{j-1}$, $\Delta^2 c_j = c_j - c_{j-1} + c_{j-2}$ and so forth for higher q. For an excellent overview of this topic, see Eilers et al. (2015) and Eilers and Marx (2010). P-splines are in the area of functional data analysis well established and an active field of research. It should be noted that P-splines use equally-spaced knots. Hence, multiple knots at both ends are not possible. Meyer (2008, 2012) developed penalized splines under constrained shape, such as monotonicity or convexity.

2.1.4 Nadaraya-Watson kernel estimator

Another idea to obtain a smooth curve is to follow a localized weighting principle. The points around a given point x should have the most influence on the fit. In order to explicitly include this, a weight function is introduced that highlights this local dependency. Suppose the estimate depends on the local points, we define

$$\hat{g}(x) = \sum_{i=1}^{n} w_i(x) y_i, \tag{2.9}$$

where the weights are given by

$$w_i(x) = \frac{\mathrm{Kern}\left(\frac{x_i - x}{h}\right)}{\sum_{r=1}^{n} \mathrm{Kern}\left(\frac{x_r - x}{h}\right)}, \tag{2.10}$$

with the kernel, Kern(u), and its bandwidth h. One commonly used kernel is the *Gaussian kernel* defined by

$$\mathrm{Kern}(u) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}u^2}. \tag{2.11}$$

2.1.5 Application

The discussed methods can be applied to the twitch.tv data, summarized in Table 2.1. The inherent granularity of the data allows two approaches. First, the underlying function is estimated based on the average number of viewers for the seven weekdays, which are displayed in Table 2.1. Second, the temporal granularity of the data set is exploited to replace each weekday's average by its 24 hourly averages for each of the 3 years. So, instead of 7 points, 504 points are used. The results are shown in Fig. 2.2, and all estimated functions are evaluated for 168 equally-spaced points between 0.5 and 7.5.

For the averages in the top panel, the non-penalized models show a similar pattern. The solid red line of the Nadaraya-Watson estimates (Gaussian kernel, $h = 2$) shows slightly less extreme values compared to the dot-dashed purple line of the regression spline ($\log(GCV) = 11.22$, $\lambda = 0$, $J = 6$). The penalized methods show almost identical shapes. The yellow dotted smoothing spline ($\log(GCV) = 9.18$, $\lambda = 0.001$, $J = 9$) follows closely the P-spline estimates. The blue dashed concave constrained P-spline

($\log(GCV) = 11.37$, $\lambda = 0.033$, $J = 6$) overlaps with the long-dashed green unconstrained P-spline ($\log(GCV) = 11.40$, $\lambda = 0.033$, $J = 6$) except the section after Sunday. Here, the concavity constraint restricts the blue dashed line compared to the small upward kink of the long-dashed green line. The mid-week peak and the higher viewer numbers on workdays are very prominent.

The bottom panel shows the smoothed fit for $n = 504$ hourly observations on an annual basis. The data is acquired as follows. First, the observations on weekday index x and the corresponding 24-hour t are converted to a common scale by $t' = x + \frac{t-12}{24}$, so, e.g., Monday at 8 p.m. is denoted by $t' = 1.33$ and $t' = 2$ represents Tuesday noon. The gray dots represent the annual average number of viewers for each t'. The number of observations for the 168 unique t' (7 weekdays, 24 hours, $7 \cdot 24 = 168$) for 3 years is $n = 3 \cdot 168 = 504$.

The non-penalized methods in the left column of Fig. 2.2 show a very similar pattern compared to the top panel when considering the y-axis scale. Thus, the data variation does not affect the Nadaraya-Watson (Gaussian kernel, $h = 2$) and regression spline ($\log(GCV) = 13.26$, $\lambda = 0$, $J = 6$). However, the penalized methods show changes. The unconstrained P-spline ($\log(GCV) = 18.92$, $\lambda = 0.011$, $J = 45$)[2] provides a less rougher fit than the smoothing spline ($\log(GCV) = 12.40$, $\lambda = 2.42e-07$, $J = 88$) but both attempt to capture the peaks of the data. The concave constrained P-spline ($\log(GCV) = 19.47$, $\lambda = 0.011$, $J = 45$) shows a similar pattern compared to the non-penalized methods but with a sharper increase at the week's beginning, a plateau around mid-week, and a sharper decrease to the week's end.

2.2 Note on testing differences between two groups

In experimental economics, almost always, researchers encounter a situation where two groups must be compared. Typically, the response variable of interest has a continuous nature. However, the visual inspection of the data set raises suspicions that it does not follow a normal distribution. Then the general advice is to use a non-parametric test. This brief note casts some doubts about this advice.

There are two competitors in this scenario – the *t-test* and the *Mann-Whitney U* (MWU) *test*. Analyzing these two tests is nothing new, as Gibbons and Chakraborti (1991) compared both for normally distributed data sets in a simulation study. However, they used very few observations and did not consider positively skewed distributions. Fay and Proschan (2010) present a detailed overview of various hypotheses for both tests from a theoretical perspective. In addition, they provide recommendations on which test is applicable for which hypothesis. In a very recent Monte Carlo study, Knief and Forstmeier (2021) investigated fitting Gaussian models to non-normal data and found that these models provide robust results despite the violation of the normality assumption. Their

[2]The data-driven methods that are implemented in the `R` routines automatically determine the knot sequence and cause the differences in J. The Js are not normalized to illustrate the parameter dependency. This section only serves for illustration and does not attempt an optimized fit.

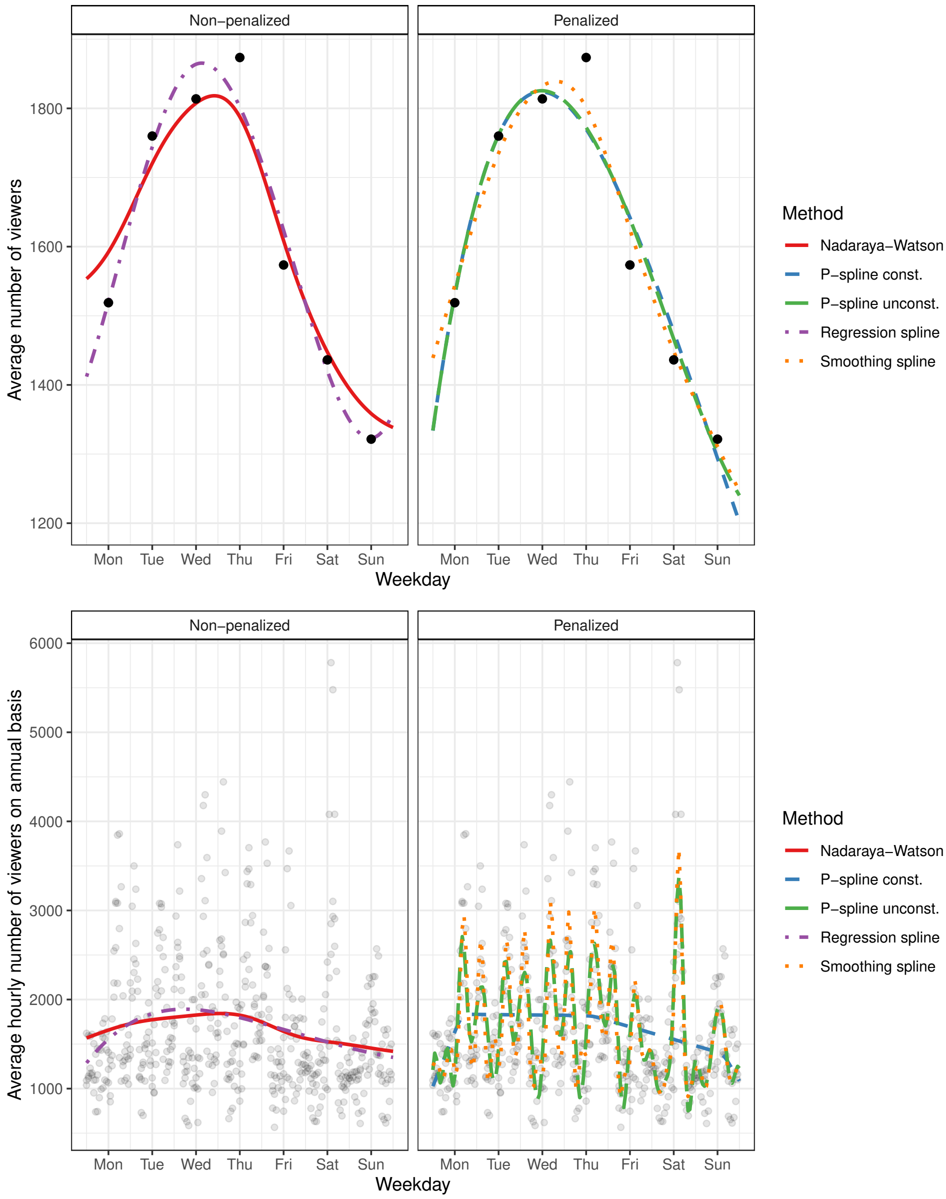


Figure 2.2. Application of the methods.

Note: The lines illustrate the application of the five methods presented in this section. The left column labeled *Non-penalized* refers to methods without a smoothing parameter λ. The right column labeled *Penalized* covers methods with a penalty term. The top panel shows the smoothed fit for the $n = 7$ points in Table 2.1 that describe the average numbers of viewers for each day. Black dots mark the observed points. The bottom panel shows the smoothed fit for $n = 504$ hourly observations on an annual basis. The gray dots represent the annual average number of viewers for each hour $t' = x + \frac{t-12}{24}$, where x is the observation's weekday index and t the corresponding 24-hour integer. The number of observations is $n = 168 \cdot 3 = 504$ for 168 unique t's (24 hours, 7 weekdays, $24 \cdot 7 = 168$) for 3 years. The line type and the coloring differentiate the methods.

literature overview presents opposing opinions on the relevance of the data's normality assumption. On the one side, it is argued that this assumption is the least important one of all assumptions; on the other hand, voices are raised that it should always be controlled for normality as its absence might lead to distortions in the inference.

2.2.1 Simulation procedure

In this note, a Monte Carlo simulation sheds more light on the performance of these two tests. The simulation follows the idea that we observe two groups, e.g., a control and treatment group, and we suspect that the treatment truly shifts the density curve of the treatment group to the right. The median is not affected, but a shift in means occurs. The simulations will highlight that the MWU cannot detect the shift. The MWU can find a shift in medians when the shape of both groups is similar, but a shift in means only if the mean collides with the median. This is the case if the distribution is symmetrical.

More formally, suppose there are independent samples from two distributions denoted by $i = 1, 2$. The test hypotheses aim to identify whether the difference in means is significant or not. Both samples are randomly drawn from distributions of the same family with an identical median (*median equality*). There is an actual shift in means so that $\mu_1 < \mu_2$. The sample size of the first group, n_1, is smaller than the one of the second group, n_2; in particular, $2n_1 = n_2$. The standard deviation should be larger for the second group $SD_1 < SD_2$.

The normal distribution is deliberately excluded from this simulation. This is to account for the following scenario: A researcher finds that the data are not normally distributed. He immediately follows the general advice and uses only non-parametric tests.

Only a few continuous probability distributions allow a parameter configuration keeping the same median and shifting the mean. Namely, the *lognormal distribution* and *Weibull distribution* fulfill the requirements. Both have applications in various fields. For example, the lognormal distribution can describe the comment length in Internet discussion fora (Sobkowicz et al., 2013), the time needed for the maintenance of engineered systems (O'Connor and Kleyner, 2011) or the stock prices in the renowned Black–Scholes model (Black and Scholes, 1973). The Weibull distribution, named after the Swedish engineer and mathematician Waloddi Weibull, can model, for example, the yield strength of steel, fractures in concrete, or wind speed distributions (Murthy et al., 2004). Both distributions and their relevant properties are summarized in Table 2.2. For each distribution, two examples of the density functions are on top of Tables 2.3 and 2.4. The figures highlight the positively skewed distributions, their medians, and means in dependence of the parameters.

The general procedure is as follows. First, the parameters of the distributions $i = 1, 2$ are determined. Distribution $i = 1$ serves as a baseline, as there is no change in the corresponding parameters that determine the density function. The parameters of distribution $i = 2$ are adjusted so that the mean and standard deviation increase, but the median remains unchanged. Second, a sample of size n_1 is drawn from the baseline $i = 1$

Table 2.2. Probability distributions and their properties.

X_i	Lognormal	Weibull
Parameters	$m > 0,$ scale, median $\sigma > 0,$ shape, sd of log	$\lambda > 0,$ scale $k > 0,$ shape
PDF, $f(x;\cdot)$	$\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(\frac{-\left(\frac{\log x}{m}\right)^2}{2\sigma^2}\right)$	$\begin{cases}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left[-\left(\frac{x}{\lambda}\right)^k\right], & x \geq 0,\\ 0, & x < 0.\end{cases}$
Mean, $\mathrm{E}\,([]\,X_i] = \mu_i$	$m\exp\left(\frac{1}{2}\sigma^2\right)$	$\lambda\Gamma\left(1+\frac{1}{k}\right)$
Variance, $\mathrm{Var}(X_i)$	$m^2\exp(\sigma^2)\,(\exp(\sigma^2)-1)$	$\lambda^2\left[\Gamma\left(1+\frac{2}{k}\right)-\left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right]$
Standard deviation, SD_i	$\sqrt{\mathrm{Var}(X_i)}$	$\sqrt{\mathrm{Var}(X_i)}$
Median, m_i	m	$\lambda\,(\log 2)^{1/k}$
		Γ is the gamma function.

Note: The table is based on Forbes (2011).

and a sample of size n_2 from distribution $i = 2$ with the adjusted parameters. Then, these two samples are compared by the t-test and the MWU test. The two reported p-values are collected. For each set of adjusted parameters, the procedure is repeated 10,000 times. The parameters are arbitrarily chosen but carefully adjusted for similarity in the moments between both distributions.

For the lognormal distribution, the baseline is given by the median $m_1 = exp(1) = e$ and $\sigma_1 = 0.5$. The median of the distribution is determined by the scale parameter m. Hence, the median $m_2 = e$ for all adjustments. So, only σ_2 varies and is given by $\sigma_2 \in \{0.5, 0.75, 1.0, 1.25, 1.5, 1.75\}$. The case of $\sigma_2 = 0.5$ serves as a check that when both distributions only differ in their sample size, the power of the test should equal the significance level.

For the Weibull distribution, the baseline is given by $\lambda_1 = 1$ and $k_1 = 1.5$. So, the median is $m_1 = (\log 2)^{1/1.5} = 0.78$. For distribution $i = 2$, given a change in λ_2, the parameter k_2 needs to adjust for the median equality $m_1 = m_2$ to hold. The parameter λ_2 is given by $\lambda_2 \in \{1, 1.25, 1.5, 1.75, 2.0, 2.25\}$. It can be easily shown that choosing k_2 by

$$k_2 = \frac{\log(\log 2)}{\log\left(\frac{\lambda_1}{\lambda_2}\right) + \frac{1}{k_1}\log(\log 2)}, \tag{2.12}$$

guarantees the median equality $m_1 = m_2$ to hold for a given λ_2 and the baseline.

Proof. Substituting the medians in the median equality by the corresponding formulas of Table 2.2 yields

$$\begin{aligned} m_2 &= m_1, \\ \lambda_2(\log 2)^{1/k_2} &= \lambda_1(\log 2)^{1/k_1}, \end{aligned}$$

$$\begin{aligned}(\log 2)^{1/k_1} &= \frac{\lambda_1}{\lambda_2}(\log 2)^{1/k_1},\\ \frac{1}{k_2}\log(\log 2) &= \log\left(\frac{\lambda_1}{\lambda_2}\right) + \frac{1}{k_1}\log(\log 2),\\ k_2 &= \frac{\log(\log 2)}{\log\left(\frac{\lambda_1}{\lambda_2}\right) + \frac{1}{k_1}\log(\log 2)}.\end{aligned}$$

□

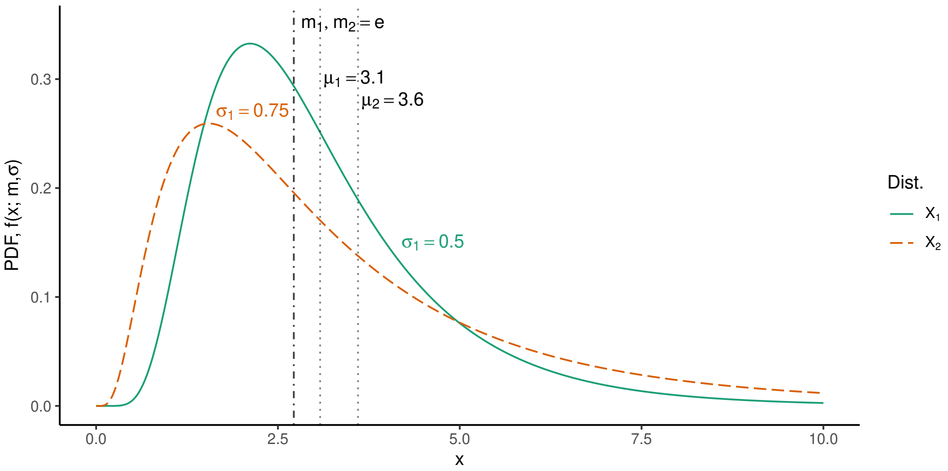
2.2.2 Results

The simulation study is based on data in violation of the normality assumption. This assumption is commonly required for the t-test. In the case of the lognormal distribution, Table 2.3 paints a clear picture that, although the data are not normal, the t-test clearly identifies the differences in means for row 5 to row 24 with a power of 46.9% for a small sample but yield more than 93.9% when the mean differences and standard deviation increase. For large samples $n_1 >= 1000$, the power reaches even higher levels. The control cases, rows 1 to 4, show that both tests only reject the null in around 5% of the cases, which is the significance level. The MWU test yields power $\leq$ 3.9% for rows 5 to 24, which is in line with the median equality. But the low power would lead to misinterpretations of the data if the MWU test is intended to detect a shift in means. It becomes very clear that despite the violation of the normality assumption, the t-test is a reasonable choice. It is not advisable to resort to the MWU test only based on the non-normal data, particularly with positively skewed data and differences in the sample sizes.

Table 2.4 confirms these results. Both tests yield in the control cases (rows 1 to 4) a power around the 5% significance level. In general, both achieve higher power for Weibull distributed data than lognormal data. Nevertheless, the t-test is close to 100%, while the MWU test has at most 19.8%. As expected, the MWU test performs better for large sample sizes than for smaller ones.

To summarize, as Knief and Forstmeier (2021) titled their paper *"Violating the normality assumption may be the lesser of two evil"* and Fay and Proschan (2010) conclude *"[t]he choice between t- and [Mann-Whitney U tests] should not be based on a test of normality."*, this note reaches the same conclusion: the t-test might be a reasonable choice even when there is non-normal data.

Table 2.3. Power analysis of lognormal distributed samples.

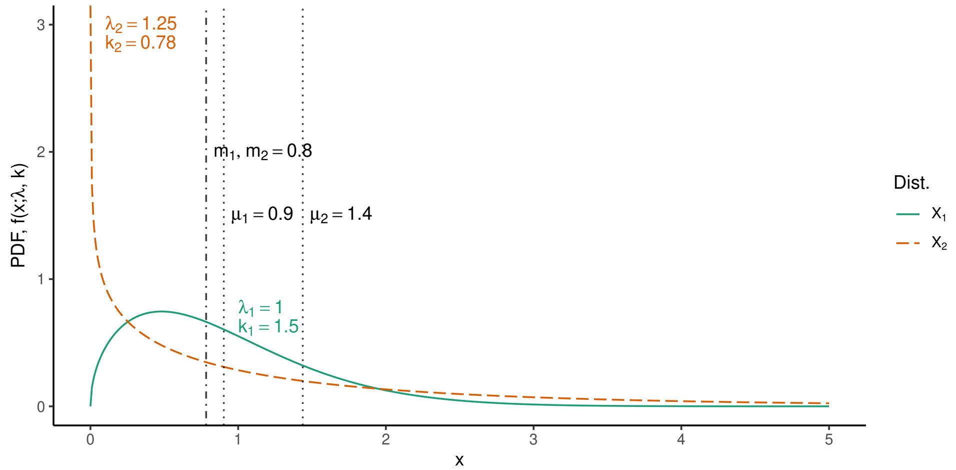


Note: The vertical lines mark by a dot-dashed line the median, and by dotted lines the means. The lines are labeled by the symbols on the right.

	n_1	n_2	$m_1 = m_2$	σ_1	σ_2	μ_1	μ_2	SD_1	SD_2	$power_t$	$power_{MWU}$
1	100	200	2.72	0.5	0.50	3.08	3.08	1.64	1.64	0.054	0.055
2	200	400	2.72	0.5	0.50	3.08	3.08	1.64	1.64	0.053	0.049
3	500	1000	2.72	0.5	0.50	3.08	3.08	1.64	1.64	0.048	0.052
4	1000	2000	2.72	0.5	0.50	3.08	3.08	1.64	1.64	0.049	0.047
5	100	200	2.72	0.5	0.75	3.08	3.60	1.64	3.13	0.469	0.037
6	200	400	2.72	0.5	0.75	3.08	3.60	1.64	3.13	0.771	0.036
7	500	1000	2.72	0.5	0.75	3.08	3.60	1.64	3.13	0.990	0.039
8	1000	2000	2.72	0.5	0.75	3.08	3.60	1.64	3.13	1.000	0.038
9	100	200	2.72	0.5	1.00	3.08	4.48	1.64	5.87	0.939	0.034
10	200	400	2.72	0.5	1.00	3.08	4.48	1.64	5.87	0.999	0.033
11	500	1000	2.72	0.5	1.00	3.08	4.48	1.64	5.87	1.000	0.033
12	1000	2000	2.72	0.5	1.00	3.08	4.48	1.64	5.87	1.000	0.032
13	100	200	2.72	0.5	1.25	3.08	5.94	1.64	11.53	0.995	0.032
14	200	400	2.72	0.5	1.25	3.08	5.94	1.64	11.53	1.000	0.030
15	500	1000	2.72	0.5	1.25	3.08	5.94	1.64	11.53	1.000	0.034
16	1000	2000	2.72	0.5	1.25	3.08	5.94	1.64	11.53	1.000	0.035
17	100	200	2.72	0.5	1.50	3.08	8.37	1.64	24.39	0.988	0.034
18	200	400	2.72	0.5	1.50	3.08	8.37	1.64	24.39	0.999	0.031
19	500	1000	2.72	0.5	1.50	3.08	8.37	1.64	24.39	1.000	0.035
20	1000	2000	2.72	0.5	1.50	3.08	8.37	1.64	24.39	1.000	0.032
21	100	200	2.72	0.5	1.75	3.08	12.57	1.64	56.74	0.969	0.034
22	200	400	2.72	0.5	1.75	3.08	12.57	1.64	56.74	0.991	0.032
23	500	1000	2.72	0.5	1.75	3.08	12.57	1.64	56.74	0.999	0.029
24	1000	2000	2.72	0.5	1.75	3.08	12.57	1.64	56.74	1.000	0.031

Note: The table summarizes the results of 24 parameter configurations. For the sample sizes, it holds that $n_2 = 2n_1$. The table is arranged by σ_2 in ascending order with increasing sample sizes. Consequently, SD_2 is increasing. The significance level is $\alpha = 0.05$. The column $power_t$ and $power_{MWU}$ summarize the fraction of 10,000 simulations where the p-value of the corresponding test was smaller than α, or in other words, the null hypothesis was correctly rejected.

Table 2.4. Power analysis of Weibull distributed samples.



Note: The vertical lines mark by a dot-dashed line the median, and by dotted lines the means. The lines are labeled by the symbols on the right.

	n_1	n_2	λ_1	λ_2	k_1	k_2	$m_1 = m_2$	μ_1	μ_2	SD_1	SD_2	$power_t$	$power_{MWU}$
1	100	200	1	1.00	1.5	1.50	0.78	0.9	0.90	0.61	0.61	0.052	0.053
2	200	400	1	1.00	1.5	1.50	0.78	0.9	0.90	0.61	0.61	0.052	0.051
3	500	1000	1	1.00	1.5	1.50	0.78	0.9	0.90	0.61	0.61	0.051	0.052
4	1000	2000	1	1.00	1.5	1.50	0.78	0.9	0.90	0.61	0.61	0.050	0.051
5	100	200	1	1.25	1.5	0.78	0.78	0.9	1.44	0.61	1.85	0.979	0.046
6	200	400	1	1.25	1.5	0.78	0.78	0.9	1.44	0.61	1.85	1.000	0.057
7	500	1000	1	1.25	1.5	0.78	0.78	0.9	1.44	0.61	1.85	1.000	0.093
8	1000	2000	1	1.25	1.5	0.78	0.78	0.9	1.44	0.61	1.85	1.000	0.169
9	100	200	1	1.50	1.5	0.56	0.78	0.9	2.46	0.61	4.68	1.000	0.045
10	200	400	1	1.50	1.5	0.56	0.78	0.9	2.46	0.61	4.68	1.000	0.061
11	500	1000	1	1.50	1.5	0.56	0.78	0.9	2.46	0.61	4.68	1.000	0.109
12	1000	2000	1	1.50	1.5	0.56	0.78	0.9	2.46	0.61	4.68	1.000	0.198
13	100	200	1	1.75	1.5	0.46	0.78	0.9	4.21	0.61	10.77	1.000	0.046
14	200	400	1	1.75	1.5	0.46	0.78	0.9	4.21	0.61	10.77	1.000	0.060
15	500	1000	1	1.75	1.5	0.46	0.78	0.9	4.21	0.61	10.77	1.000	0.100
16	1000	2000	1	1.75	1.5	0.46	0.78	0.9	4.21	0.61	10.77	1.000	0.185
17	100	200	1	2.00	1.5	0.39	0.78	0.9	7.09	0.61	23.13	0.998	0.044
18	200	400	1	2.00	1.5	0.39	0.78	0.9	7.09	0.61	23.13	1.000	0.059
19	500	1000	1	2.00	1.5	0.39	0.78	0.9	7.09	0.61	23.13	1.000	0.101
20	1000	2000	1	2.00	1.5	0.39	0.78	0.9	7.09	0.61	23.13	1.000	0.174
21	100	200	1	2.25	1.5	0.35	0.78	0.9	11.62	0.61	46.84	0.991	0.047
22	200	400	1	2.25	1.5	0.35	0.78	0.9	11.62	0.61	46.84	1.000	0.056
23	500	1000	1	2.25	1.5	0.35	0.78	0.9	11.62	0.61	46.84	1.000	0.095
24	1000	2000	1	2.25	1.5	0.35	0.78	0.9	11.62	0.61	46.84	1.000	0.159

Note: The table summarizes the results of 24 parameter configurations. For the sample sizes, it holds that $n_2 = 2n_1$. The table is arranged by λ_2 in ascending order with increasing sample sizes. Consequently, SD_2 is increasing. The significance level is $\alpha = 0.05$. The column $power_t$ and $power_{MWU}$ summarize the fraction of 10,000 simulations where the p-value of the corresponding test was smaller than α, or in other words, the null hypothesis was correctly rejected.

Chapter 3

Neural Functional Time Series Forecasting

Electricity is a unique commodity. It cannot economically be stored, and consequently, a permanent balance between electricity generation and demand is an essential requirement for the market. Suppliers and purchasers are heavily affected by changes in business activities, weather conditions, technology developments, and the political situation. As a result, the prices and load demands show abrupt, generally unanticipated spikes, which makes forecasting a challenging discipline for market participants.

The recent developments of the energy market have led to a stark increase in this particular market's relevance. The German government made a complete turn-over from the nuclear power plants' prolongation to a complete phase-out from nuclear energy until 2022. In light of climate change, greenhouse gas emissions, and air pollution, the demand for low-carbon electricity sources emerges naturally. However, most of these sources introduce frequency variations through their unsteady availability, which the *transmission system operators* (TSOs) have to balance following the load-frequency control concept. To compensate for spikes or gaps in demand and supply, the TSOs require balancing power[3], which they procure in auctions on the balancing market. Accurate predictions are, therefore, paramount for well-informed decision-making.

This paper outlines a new approach to predict supply curves based on historical data by a highly specialized neural network. Such supply curves typically emerge from the auctions in the German balancing market, representing an essential institutional arrangement for the energy market to guarantee energy security. However, these curves can also be observed for other products of wholesale markets, or storage capacity planning, or generally in sales and many other areas, making the approach generally well suited for forecasting tasks.

A small stylized example can illustrate three things in one sweep: the observed data structure, its challenging characteristics, and this paper's approach. In procurement auctions, a single buyer interacts with multiple sellers who offer bids of a quantity x for a

[3]On the official website, the TSOs define control reserves as balancing capacity and balancing energy. The literature refers more often to the term balancing power (Ocker et al., 2018; van der Veen and Hakvoort, 2016).

price p. The buyer will accept the lowest price and accept higher prices in increasing order until his demand is met (see, e.g., the *merit-order*). So, for each auction exists a natural monotone curve when the buyers' procured quantities are mapped to the ordered prices in the form of a step function. It is easy to see that these curves are observed on irregularly spaced intervals caused by variation in the bid sizes and not truncated at random but by the buyer's demand. With a simple adjustment, the proposed model can deal with these step functions, while classical approaches require interpolation, smoothing, or truncation. For illustration, Fig. 3.1 shows two stylized curves of subsequent periods. The first two blue triangles represent that a seller made an offer with quantity three ($x = 5 - 2 = 3$) for a price of four ($p = 4$).

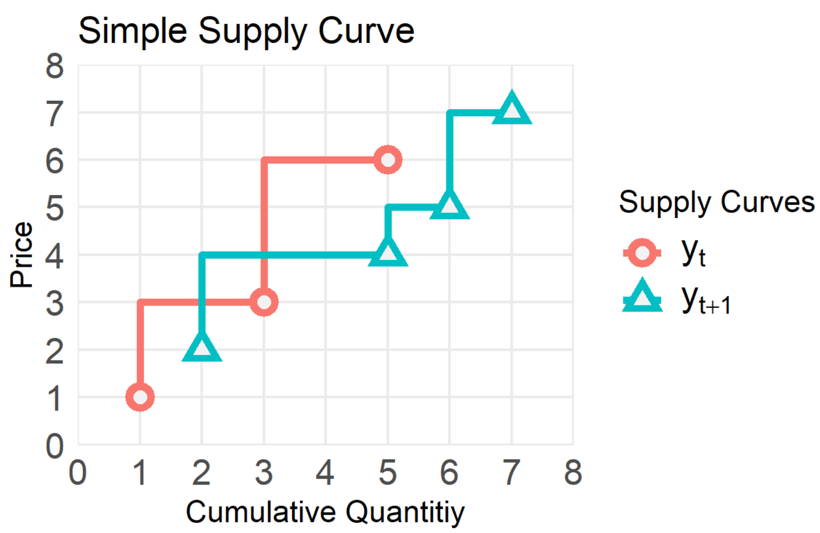


Figure 3.1. Stylized supply curve example.

The underlying key concept is that such curves represent next-generation functional data (Wang et al., 2016), namely a *functional time series*, and at the same time, they also fall in the category of *sequences* in the sense of the machine learning literature (Goodfellow et al., 2016). If this sequence is a sentence for a translation task, the sequence is governed by the rules and principles of the language, while the functional time series is a set of observations determined by one generating process. There are well-known methods from the artificial intelligence field for sequence-to-sequence (*seq2seq*) prediction. The recurrent neural network or RNN (Rumelhart et al., 1986) is specialized for sequential data and can deal with multidimensional data; for a detailed introduction, see Graves (2012). By looping through the sequence's elements, it shares relevant information across each iteration allowing the network to remember specific features of the sequence. In the field of neural machine translation, Cho et al. (2014a), Kalchbrenner and Blunsom (2013), and Sutskever et al. (2014) simultaneously developed based on RNNs the *Encoder-Decoder* architecture, which is renowned and the foundation of many practical applications. Its original aim was to translate one sentence of an input language to a target language. Since sentences with the same meaning but in different languages frequently have varying sequence lengths, this architecture is well suited for seq2seq predictions of different lengths, especially for the case of supply curves with different numbers of bids.

The proposed model is based on an RNN with encoder-decoder architecture with an attention mechanism and an imposed autoregressive input data structure (RNAA) to tackle the challenges of forecasting supply curves. An intuitive explanation of the working procedure of the RNAA is as follows. Suppose, returning to the stylized example in Fig. 3.1, we would like to predict the future supply curve (blue curve) based on the currently observed supply curve (red curve). Initially, the RNAA tries to comprehend the relations within the observed supply curve by starting at the first point and moving along the sequence while storing important information in a compressed format. After that, it evaluates the relationship between the complete observed supply curve and the first point of the future supply curve with this compressed knowledge. Then it moves along the sequence of the future supply curve from point to point. Moreover, simultaneously to the previous step, the attention mechanism allows the RNAA to attach weights to the compressed knowledge to autonomously select relevant information of the observed curve for each future curve point.

The performance assessment of the RNAA consists of three steps, where the German balancing market serves as the data source for the supply curves. The analysis focuses mainly on the price dimension, but the raw data require processing. In a first step, taking averages transforms the curves into a univariate time series which makes the comparison with classical methods possible. The next step requires a twofold normalization of the supply curves to a common domain on regularly spaced intervals to apply the functional time series approach, which a joint truncation and smoothing procedure achieves. In the last step, all models aim to predict the supply curves as observed without normalization. The RNAA can simultaneously forecast multi-entry bids consisting of price, bid size, and covariates by simply adjusting the input data structure. It illustrates the potential of the RNAA since approaches from the first and the second step cannot do this without relying on multiple models or more complex combined approaches. Therefore, in the third step, the RNAA's predictions are compared with the non-truncated supply curve, while the other methods are compared with the truncated ones. In all steps, the RNAA shows promising performance compared to the benchmark models.

The article is organized as follows. Section 3.1 gives an overview of the related literature. Section 3.2 introduces the RNAA model in more detail, and Section 3.3 briefly presents relevant benchmark methods and forecast assessment metrics. Section 3.4 reveals details on the data source, namely the German balancing market and the data processing. In general, neural networks are highly dependent on hyperparameters that govern their learning, and Section 3.5 discusses their optimization in the RNAA environment. Section 3.6 illustrates the performance of the RNAA and presents a methodology to visualize how the RNAA perceives the data during training. Section 3.7 summarizes some notes on the implementation, and Section 3.8 concludes.

3.1 Related literature

This project contributes to multiple strands in the literature. It provides a methodology for electricity price forecasting when the aim is to predict supply curves, a novel application of machine translation methods, and a new rich tidy data set for applications of functional data analysis describing the supplier behavior of the German balancing market.

The literature for electricity price forecasting has rapidly grown within the last two decades (Nowotarski and Weron, 2018). A large body of work exists utilizing methods that range from time series analysis over machine learning algorithms to combined techniques. To pool these approaches of various disciplines, the Global Energy Forecasting Competition was established, which shows the relevance for industry and research alike (Hong et al., 2014, 2016, 2019).

There are multiple surveys of the *energy price forecasting* (EPF) literature. Aggarwal et al. (2009) categorize the approaches in game theoretical models, time series models, and simulation models and find that no category systematically outperforms the others. The overview by Chan et al. (2012), discussing methodologies in the context of smart grids, additionally introduces the *functional principal component analysis* (FPCA) models, which serve as the basis for their robust FPCA approach. Weron (2014) provides a systematic and extensive review of the EPF literature and the proposed models and methods. The author discusses statistical and computational intelligence models separately and concludes with a look into the future of EPF, which covers the correct covariate choice, the use of combined models, and a call for a standardized testing procedure, among other topics. Nowotarski and Weron (2018) continue with a focus on probabilistic forecasting and find that the relevance of machine learning methods in the literature on EPF sharply increased in recent years. As they call for closer cooperation between electrical engineering focused on computational intelligence and econometric approaches, the article at hand presents a symbiosis where a neural network allows to overcome the limitations of classical time series approaches without imposing restrictions on the characteristics of supply curves.

Ziel and Steinert (2018) provide an update of the review mentioned above with a focus on medium- and long-term price forecasting. In addition, they extend the X model, which was introduced in Ziel and Steinert (2016), for a long-term horizon. It models the purchase and sales curves of an auction separately, and the interaction of both curves represents the electricity price. Shah and Lisi (2020) provide a similar idea in the world of functional time series and show its performance in an application to the Italian electricity market. Ziel and Steinert (2018) report that only two recent forecasting projects use data from the German market in their applications. Another example for the German electricity market can be found in Gianfreda et al. (2020), who compare various univariate and multivariate time series models considering renewable energy sources for predicting hourly day-ahead electricity prices. They find that multivariate Bayesian models with exogenous variables lead to improvements in all markets. Overall, EPF is an active research area with a long tradition. The present paper contributes to this field a novel approach and application to the German electricity market that is rarely considered in the literature.

Neural networks are among the top research fields in the machine learning literature (Glauner et al., 2017). Neural networks consist of at least one input, output, and possibly multiple hidden layers, where each contains several cells or nodes. Typical data sources for time series forecasting are the stock and electricity market due to the large quantity of available information. RNNs are well suited for these large data sets, but the seq2seq predictions require a seemingly univariate time series to be transformed to sequences. In the context of electricity prices, the approaches cleverly exploit the more granular underlying time structure. For example, if hourly prices are observed, 24 of them are bundled to generate one daily observation. See, e.g., Gonzalez et al. (2018) for an application using functional time series analysis. Bundling data this way allows to obtain sequence with a fixed window of 24 observations. This collection window can be moved step-wise ahead. This transformation is henceforth referred to as *Window approach*. An example using this paper's data will be given in Section 3.4.3. For an extensive benchmark study of EPF for the Belgian market, see Lago et al. (2018), and for a survey of time series forecasting applications with neural networks, see Lim and Zohren (2021).

The TSOs operate the balancing market where pre-qualified suppliers compete to provide balancing power products, which are used to smooth frequency deviations in the power grid to guarantee system stability. In particular, the TSOs periodically call for tenders of these products in a multi-unit procurement auction (Zweifel et al., 2017, Ch. 13; Weron, 2006). With the large national differences in these markets of European countries, the discussion of an optimal balancing market design has been initiated in the work of van der Veen and Hakvoort (2016), Ocker et al. (2016), and Vandezande et al. (2010) among others. The recent literature has focused on theoretical analyses with distributed energy resources integration (e.g. Borne et al., 2018; Poplavskaya and de Vries, 2019) or strategic bidding behavior of market participants (e.g. Campos et al., 2016; Mazzi et al., 2018; Poplavskaya et al., 2020). For the German case, Müsgens et al. (2014), discuss the market design and its economic fundamentals, and Ocker et al. (2018) provide a game-theoretical analysis of the auction format with real-data applications. Liebl and Rameseder (2019) propose a novel functional data estimation method for the mean and covariance function under violation of the missing at random assumption, which the supply curves of the German balancing market exhibit. This paper also looks at the supply curves of this peculiar market but takes the path to develop an improved prediction method.

The balancing markets exhibit high price volatilities, leading to little empirical work in the forecasting discipline for this market in general (Weron and Ziel, 2018) and even less for the German market. There are some works on the Nordic market, where Klæboe et al. (2015) compare various time-series model-based forecast techniques for the Norwegian market and Boomsma et al. (2014) present a theoretical model of bidding strategies with empirical analysis of spot and balancing prices. Olsson and Soder (2008) propose a model for balancing power price scenarios for scenario trees based on SARIMA and Markov processes for the Nordic power market. For the UK, Lucas et al. (2020) found that using the loss of load probability for forecasting the balancing market price via XGBoost, Gradient Boosting, and Random Forest leads to more precise predictions. However, the task to forecast the supply curves of the German balancing market remains an open task.

This paper tackles this task and presents a new prediction method, taking into account the unique shape of supply curves. For this purpose, neural networks are used, which seem to be made for this task. Moreover, these are flexible enough to meet future challenges, such as those posed by the increasing penetration of the market by renewable energies and distributed energy resources (Hirth and Ziegenhagen, 2015; Zweifel et al., 2017). To keep the scope of this paper to reasonable limits, the application focuses on the positive automatic frequency restoration reserve (aFRR) due to its relevance in the literature (Borne et al., 2018; Ocker et al., 2018; Poplavskaya and de Vries, 2019).

3.2 Methodology

The RNAA follows Bahdanau et al. (2015) who developed a model that learns to align and translate simultaneously. Let $\mathbf{x} = \left(x^{(1)}, \ldots, x^{(n_x)}\right)$ be an input sequence of length n_x, where $x^{(i)}$ denotes the entry at position i within the sequence.[4] The superscript i is referred to as *step*. Additionally, let $\mathbf{y} = \left(y^{(1)}, \ldots, y^{(n_y)}\right)$ be a target sequence of length n_y, where $y^{(i)}$ denotes the entry at step i.[5]

In order to deal with varying sequence lengths, the encoder-decoder architecture consists of two stacked RNNs, where the first is labeled *Encoder* and the second *Decoder*.

The Decoder of the RNAA operates on the L^2 norm loss function, i.e., the mean squared error, and characterizes the model as a regression approach. For the case of two sequences $\mathbf{x}$ and $\mathbf{y}$, the loss function is given by

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = n_y^{-1} \sum_{i=1}^{n_y} \left(y^{(i)} - \hat{y}^{(i)}\right)^2, \tag{3.1}$$

where

$$\hat{y}^{(i)} = g\left(y^{(1)}, \ldots, y^{(i-1)}, \mathbf{x}\right) = g\left(y^{(i-1)}, s_d^{(i-1)}, c^{(i)}\right), \tag{3.2}$$

and $g(\cdot)$ is a nonlinear, potentially multi-layered, function, where $s_d^{(i-1)}$ is the hidden state of the Decoder collecting $y^{(1)}, \ldots, y^{(i-2)}$, and the attention mechanism of the RNAA provides the context, $c^{(i)}$, which summarizes the sequence $\mathbf{x}$. Equation (3.2) shows that the ground truth of the previous step, $y^{(i-1)}$, is fed into the network as input to predict

[4] The author is well aware that i is commonly replaced by t in the literature, and the position is referred to as the *time-step*. However, in time-series forecasting, it is more natural to distinguish different periods by t. Hence, t is used to separate curves and i for positions within the curve.

[5] To introduce the method, the entries of $\mathbf{x}$ and $\mathbf{y}$ are univariate. The extension to the multivariate case is straightforward but has been omitted for clarity. For example, consider the red curve in Fig. 3.1 and its observed points $(1,1),(3,3),(5,6)$. In the univariate case, $\mathbf{x} = (1,3,6)$ and in the multivariate case $\mathbf{x}' = \left(\begin{smallmatrix}1 & 3 & 5\\ 1 & 3 & 6\end{smallmatrix}\right)$, which should be followed by additional adjustments of relevant equations.

the next step. This design feature is called *teacher forcing* (Williams and Zipser, 1989). It allows the RNAA to adjust its prediction during training and leads to a higher accuracy by staying close to the ground truth.

The network architect chooses key elements, such as the layer types, hyperparameters, and the underlying design, generally with the application and performance in mind. The presented architecture of the RNAA is adjusted for the application and based on various simulations. The RNAA's Encoder is based on a *long short-term memory* (LSTM) cell (Hochreiter and Schmidhuber, 1997) and the Decoder on a *gated recurrent unit* (GRU) cell (Cho et al., 2014b). Hence, from a theoretical perspective, the RNAA combines the benefits from both: the LSTM allows for a long memory, and the GRU is computationally more efficient than an LSTM since it has one less gate. Besides, the attention mechanism of Bahdanau et al. (2015) is originally also based on the GRU. Both consist of multiple *gates*, whose roles are explained in the next part, and, given their similarities, address the problem of vanishing gradients (Pascanu et al., 2013). Instead of a detailed formal description of the RNAA, which Appendix A.1 presents, a process-oriented description was chosen. It follows Fig. 3.2 from the bottom left to the top left corner. The right-sided gray rectangles provide more details on the Encoder and Decoder.

The Encoder reads the input sequence, $\mathbf{x}$, step-by-step and compresses for each step the relevant information in a fixed-length output vector, $\mathbf{h}$ (*annotations*), of length n_x. The length of $\mathbf{x}$ might vary from sequence to sequence. For each step of the input sequence, $x^{(i)}$, and its own previous output, $h^{(i-1)}$, the LSTM structure within the Encoder updates its internal cell state by partly forgetting its current state, $s^{(i-1)}$, (*forget gate*) and including relevant information of the input and previous output in the form of a state candidate, $\tilde{s}^{(i)}$, (*input gate*). The *output gate* governs the relationship between its new internal state, $s^{(i)}$, and output, $h^{(i)}$. This process is illustrated for the Encoder in the bottom-right corner of Fig. 3.2.

The Decoder receives as input the ground truth, $y^{(i-1)}$, the current *context*, $c^{(i)}$, and its own previous cell state, $s_d^{(i-1)}$. The context is provided by the attention mechanism that consists of a separate but very simple neural network.[6] The attention mechanism is illustrated by the dark-shaded hexagon. The network, NN, scores for each step the relevance of the annotations, $\mathbf{h}$, with the previous state, $s_d^{(i-1)}$, and summarizes the information about the annotations in the attention weight vector, α. The sum of the attention-weighted annotations depict by the octagon, $\sum |x$, is the context given to the Decoder. The GRU architecture within the Decoder updates its internal state based on two inputs, the previous state and an update candidate, $\tilde{s}_d^{(i)}$, adjusted by the *reset gate*. The *update gate* determines the share of each part for the state update. Additionally, the Decoder receives the previous step of the target sequence, $y^{(i-1)}$, as input in order to learn the sequence lengths (teacher forcing). This process is illustrated for the Decoder in the top-right corner of Fig. 3.2.

[6] In particular, the attention mechanism is a feedforward neural network with one layer and softmax activation function that is jointly trained. Bahdanau et al. (2015) originally labeled it *alignment model*.

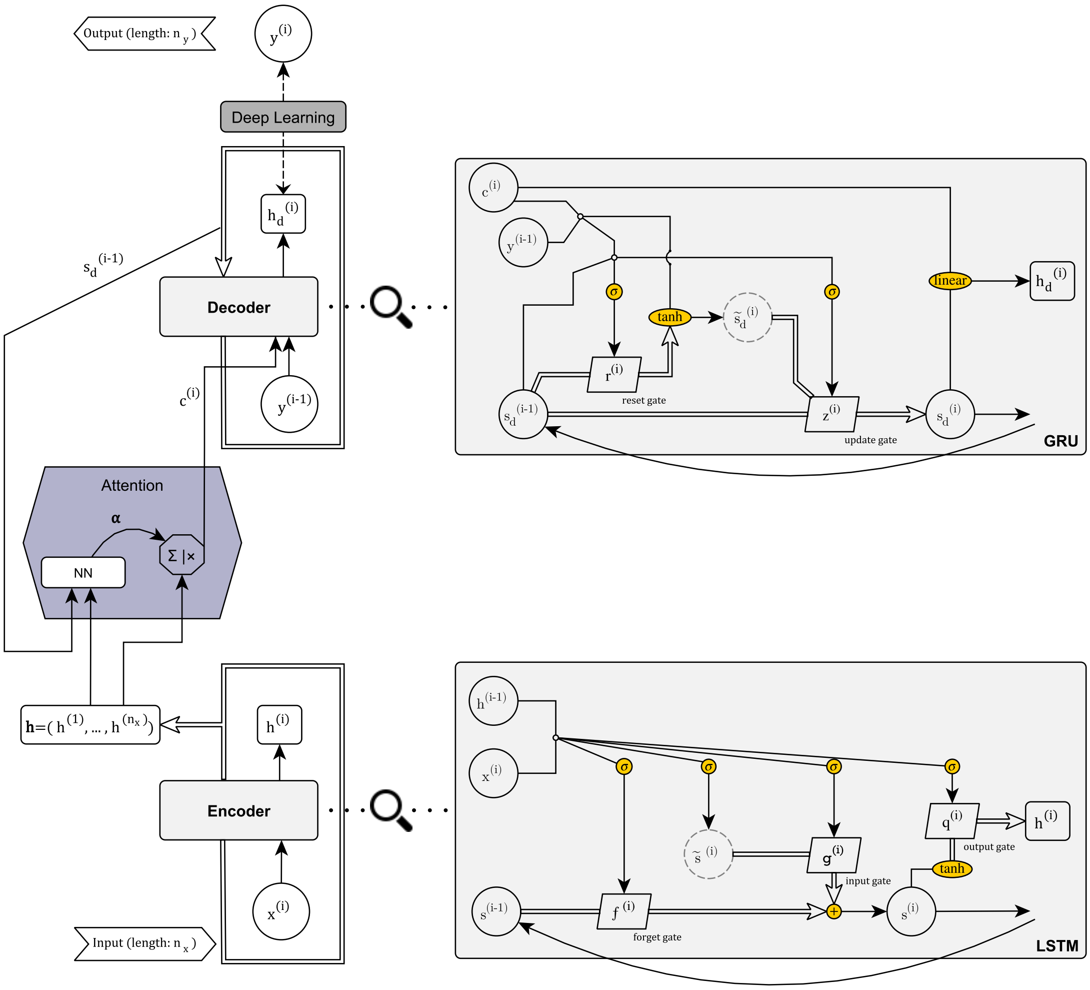


Figure 3.2. The RNAA - an encoder-decoder architecture with attention mechanism and deep learning.

Note: The Encoder incorporates an LSTM and the Decoder a GRU cell. The input and output sequences vary in their length. Two-lined closed arrows represent self-connecting loops, while single-lined symbolize an information transfer. Connecting arrows show the dependency between the figure's parts. NN is a simple neural network. Deep Learning collects optional dense layers for the deep learning structure. The double-headed arrow with a dashed line symbolizes output evaluation against actual observation. The dotted lines with loupes mark detailed representations of the Encoder and Decoder. Passing through a gate is equivalent to being multiplied by its current value. Small circles on solid lines symbolize collecting information, such as states, inputs, and outputs. The functions $\sigma, tanh$, and *linear* are defined in Appendix A.1.

So the model learns to match an input sequence, x, to a target sequence, y, and their within-sequential dependencies while autonomously deciding which part of the encoded input sequence should be paid attention to under teacher forcing.

The recent success of deep learning structures (Lago et al., 2018; Lim and Zohren, 2021) is the reason why also the RNAA allows an optional deepening of its structure. In particular, the RNAA allows a dynamically determined number of additional dense layers after the encoder-decoder architecture. Dense layers represent the simplest form of a hidden layer. In Fig. 3.2, these layers are collected as *Deep Learning.* They are placed centrally on the dashed line with double-headed arrows. If no deep learning structure is required, the Encoder-Decoder output simply passes unchanged through and is evaluated against the observation $y^{(i)}$.

The Encoder-Decoder originates from the field of neural machine translation and has proven to successfully handle the translation task, but Cho et al. (2014b) report that the performance dramatically drops when the sequence length is increased. The solution was the introduction of attention in the field of *natural language processing* (NLP); for a systematic overview, see Galassi et al. (2020) and the examples mentioned therein. The attention mechanism was recently included in one of the most famous machine learning libraries, `TensorFlow`, during its upgrade to the next version. It clearly shows the relevance of attention for state-of-the-art applications. The applications of attention are not limited to NLP (Chaudhari et al., 2020), and for an illustrative example in the field of machine vision, consider Wojna et al. (2019).

The method was originally developed for machine translation based on the cross-entropy loss function for distributions over specific vocabularies, so several adjustments had to be made to the original design for time series forecasting. The loss function was changed to the well-known MSE given by Eq. (3.1) and the usual classification output function of the Decoder, softmax, was replaced by a linear function, i.e., *as-is* output. In NLP, the number of features corresponds to the words in an ex-ante defined vocabulary, usually a few thousand. In the RNAA, this number drastically decreased to either 1, 2, or 3. The input sequence $\mathbf{x}$ was replaced by the first lag of the target sequence $\mathbf{y}$. A formal description requires the introduction of time indices. So, let the input be given by

$$\mathbf{x} = \mathbf{y_t} = \left(y_t^{(1)}, \ldots, y_t^{(n_t)}\right), \tag{3.3}$$

and the target given by

$$\mathbf{y} = \mathbf{y}_{t+1} = \left(y_{t+1}^{(1)}, \ldots, y_{t+1}^{(n_{t+1})}\right), \tag{3.4}$$

for the periods $t = 1, \ldots, T$. The encoder-decoder architecture inspired by the challenges of language translation can handle varying sequence lengths, which is shown by the different sequence lengths of each period, i.e., n_t and n_{t+1}. The imposed data structure in Eq. (3.3) and Eq. (3.4) is referred to as *autoregressive input data structure* throughout this article.

Sequential data are generated by natural phenomenons such as speaking, observing and handwriting. The sequences are governed by common rules such as grammar, semantics and meaning. The German balancing market is one potential source for economic sequences. However, these specific data are incidental truncated. This market is a good microeconomic example where supply must meet demand. The demand announced by the TSO cuts off the supply curve at the interception point, and only the accepted offers are publicly available. So, any estimation procedure assuming random samples might not be applicable, and sample correction methods require more information about the auction participants, or in other words, the missing-at-random assumption is violated. For an appropriate benchmark study, the models presented in the next section require data that are preprocessed to deal with characteristics of the supply curve.

3.3 Benchmark models

The performance of the RNAA is assessed in three steps, where step-by-step, the data are less processed for the methodological requirements. The first step is the *univariate case*, the second step is the *functional case*, and the third step is the *supply curve case*. In the last case, the supply curves are fed to the network nearly as recorded and published. The following section briefly introduces models for univariate and functional time series, and simple NN approaches. These models serve as benchmark methods for the RNAA in these cases. Since each method can handle only specific data formats, Section 3.4.3 explains the procedure that transforms the observed supply curves to the required format.

3.3.1 Univariate model

In order to evaluate the forecasting performance in the univariate time series case, the RNAA is compared to an *autoregressive moving average model with generalized autoregressive conditional heteroskedasticity* (ARMA-GARCH). The ARIMA class of model is commonly used in the literature of electricity price forecasting (see, for example, Conejo et al., 2005; Crespo Cuaresma et al., 2004; Weron and Misiorek, 2008), the GARCH models was applied to predict day-ahead prices in Spain and California by Garcia et al. (2005), and the ARMA-GARCH model were used to model mean and volatility of the electricity price of the New England market in Liu and Shi (2013).

The model ARMA(p, q)-GARCH(r, s) is given by

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t, \tag{3.5}$$

where ϕ_i is the ith autoregressive coefficient, and θ_i is the ith moving average coefficient. The error term ε_t is defined as

$$\varepsilon_t = \sigma_t z_t, \tag{3.6}$$

$$\sigma_t^2 = w + \sum_{j=1}^{r} \alpha_j \varepsilon_{t-j}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2, \tag{3.7}$$

with σ_t^2 denoting the conditional variance and w the intercept. So, ε_t follows a GARCH(r, s) process and z_t is i.i.d. with zero mean and variance of unity Bollerslev (1986). The Bayesian information criterion (BIC) and the goodness-of-fit test of Palm (1996) show that the student t-distribution firstly presented in (Bollerslev, 1987) in its skew version (Fernández and Steel, 1998) is best suited for fitting the standardized innovations in this application to the data.

It is good practice in forecasting studies to include the so-called *Naive* method. Under this method, the forecast for a period is the most recent past observation. In other words, an observation today might be a good forecast for tomorrow under certain conditions. This method is also available for the other cases but not explicitly discussed as its implementation is generally straightforward.

3.3.2 Functional time series models and neural networks

The functional time series approaches require smoothed supply curves, which are evaluated at a fixed number of points shared by each curve. This means that compared to the univariate case, more information held by the supply curves is usable but limited by the domain all supply curves share. This information gain leads to sequential data structures, allowing simple neural network approaches to be applied. The following briefly presents these benchmark methods with only the most necessary assumptions for readability.

A functional time series $(Y_t,\ t \in \mathbb{Z})$ is a sequence of curves following Hörmann and Kokoszka (2012). It is assumed that each curve, Y_t, is an element of the Hilbert space $H = L^2([0,1])$ equipped with the inner product $\langle f, g\rangle = \int_0^1 f(v)g(v)\,dv$ and is a square integrable function satisfying $\|Y_t\|^2 = \int_0^1 Y_t^2(v)\,dv < \infty$.

The *functional autoregressive* (FAR) process is studied in the monograph Bosq (2000) and the FAR(1) model is given by

$$Y_t = \Psi(Y_{t-1}) + \xi_t, \tag{3.8}$$

where ξ_t is a sequence of i.i.d. mean zero errors in H and the bounded linear operator Ψ satisfies the conditions that a unique strictly stationary causal solution for Eq. (3.8) exists. The one-step ahead prediction of a FAR(1) can be obtained by

$$\hat{Y}_{t+1} = \hat{\Psi}_d\left(Y_t\right),$$

where $\hat{\Psi}_d$ is the estimator of Ψ based on the first d most important *empirical functional principal components* (EFPCs) of the sample covariance operator. Similar to the well-known univariate AR(1) case, this estimator can be found based on the Yule-Walker equations in their functional version. Details can be found in Hörmann and Kokoszka (2012, Sec. 3.2) or in Aue et al. (2015, Sec. 3.1), where the FAR(1) serves as a benchmark for the proposed method. Since $H = L^2$, the estimator can be referred to as *Estimated Kernel* in the sense of Didericksen et al. (2012). They report that this method provides almost perfect predictions, i.e., in their simulation study, it achieved comparable performance to a method based on the true Ψ representing the theoretical but hypothetical best approach.

This is in line with the previous results of Besse et al. (2000). The associated forecast error metrics of this model are labeled *Bosq*.[7]

Aue et al. (2015) propose a new prediction methodology consisting of three steps that are intuitively appealing and utilize existing methodology with already developed implementations. It proves to be competitive or superior in performance in simulations and in an application to pollution data compared to the Bosq procedure. Their methodology consists of the following. Similar to Bosq, the first step starts with the estimation and selection of the d EFPCs by, e.g., a fraction of the data variation that needs to be explained. Typically, d is much smaller than the number of observations, often a single-digit number. With the d EPFCs, the *functional principal component* (FPC) scores are calculated. Then, in the second step, applying multivariate prediction techniques on these scores allows obtaining their forecasts. In the third and last step, these forecasts are retransformed to curves by a truncated Karhunen-Loéve representation. For more details and extensions, e.g., FAR(p) or the inclusion of covariates, refer to Aue et al. (2015).

Shang (2013) provides the `R` package `ftsa` and its current version implements among other useful tools for functional time series analysis the procedures of Aue et al. (2015) and Klepsch et al. (2017). To relate to functional time series analysis and because the approach is based on multivariate techniques, the results from this approach are referred to as *FTSA - Multi*. Earlier, Hyndman and Shang (2009) and Hyndman and Shahid Ullah (2007) proposed a similar procedure but based on univariate techniques. Hence, their approach is listed as *FTSA - Uni*. The number of components d is chosen by the multiple testing procedure of Kokoszka and Reimherr (2013).

Lastly, the benchmark study also covers three simple *neural network* (NN) architectures that are commonly used in time series forecasting. They are implemented in their single-shot version, i.e., the whole curve is predicted in one step. The *Dense* model consists of a dense layer, i.e., a densely connected neural network layer. The *CONV* model extends the Dense model by a 1D convolution layer added before the dense layer. Convolutional networks were firstly introduced in LeCun (1989). The last model consists of an *LSTM* layer followed by a dense layer. All benchmark methods are summarized in Table 3.1.

3.3.3 Forecast assessment

The performance is evaluated by a pseudo-out-of-sample scheme commonly found in econometrics, the forecasting literature, and studies in the machine learning literature. For this purpose, the models have no access to S observations during the estimation or training. That means the size of the training set is $\iota = T - S$, and the size of the test set is S. In order to assess the prediction quality, two cases must be distinguished. In the first case, the observed curve and the prediction must have the same number of points.

[7] The `R` routine for this method was gratefully received from Alexander Aue and is based on the `fda` package.

Table 3.1. Overview of benchmark methods and short descriptions.

Group	Label	Description
Univariate	*Naive*	Repetition of past observations
	ARMA-GARCH	Model by Bollerslev (1986)
Functional	*Naive*	Repetition of past smoothed supply curves
	Bosq	FAR(1) forecast in Bosq (2000)
	FTSA - Uni	Method by Hyndman and Shang (2009) and Hyndman and Shahid Ullah (2007) based on univariate techniques
	FTSA - Multi	Methods by Aue et al. (2015) and Klepsch et al. (2017), implementation by Shang (2013) based on multivariate techniques
Neural networks	*Dense*	Neural network of one densely-connected layer
	CONV	Convolutional neural network, LeCun (1989)
	LSTM	Long Short-Term Memory neural network, Hochreiter and Schmidhuber (1997)

Note: The column *Group* collects methods from the same area and also hints towards the required data structure. The column *Label* refers to the used abbreviation in the results section. The Naive approach for neural networks is omitted because it is conceptually identical to the functional case.

Both the univariate and FTSA approaches meet this requirement.[8] Then, a point-by-point comparison is possible, and the well-known *mean squared error* (MSE) and the *mean absolute error* (MAE) can be used. The second case requires a measure that can compare two functions and is thus based on integration. The measure is the *mean squared area between curves* (MSABC) and will be presented in the second part of this section.

MSE and MAE The point-by-point evaluation metrics are the MSE and the MAE. For some univariate forecast $\hat{y}_{\iota+h}$ for $h = 1, ..., S$ and some univariate observation $y_{\iota+h}$, these metrics are defined by

$$MSE = S^{-1} \sum_{h=1}^{S} \left(\hat{y}_{\iota+h} - y_{\iota+h}\right)^2 , \tag{3.9}$$

$$MAE = S^{-1} \sum_{h=1}^{S} \left|\hat{y}_{\iota+h} - y_{\iota+h}\right| . \tag{3.10}$$

It is easier to introduce the two forecasting principles used in this paper by staying briefly in the univariate world. The first principle consists of a 1-step ahead forecast based on the observation from the previous period. It can be formally described by

$$\hat{y}_{\iota+h|\iota+h-1}, \tag{3.11}$$

[8]This might be surprising at first. However, the standard FTSA procedure is that the observed curves are first smoothed and then evaluated at the same points. Thus, there is an equal number of points for smoothed observed curves and the predictions based on them.

where the basis of the forecasts rolls through the test set, period by period, for $h = 1, \dots, S$. Hence, forecasts using this procedure are marked by the label *Rolling*.

The second uses its own forecast for $h > 1$ to predict the next period and is defined as

$$\hat{y}_{\iota+h|\iota}. \tag{3.12}$$

Note the missing h in the condition in the subscript, so S-steps ahead are generated based on the training set. Forecasts under this principle are labeled *Ahead*. Naturally, the forecasts under both principles are identical for $h = 1$.

As an example, consider the *Naive* forecasting method. Because the method is optimal when data follow a random walk without drift, the forecasts are also called random walk forecasts. For Rolling they can be obtained by collecting $y_{\iota+h-1}$, for $h = 1, \dots, S$ and for Ahead by repeating y_ι, for S times. A selected method's forecasts replace $\hat{y}_{\iota+h}$ in Eq. (3.9) and Eq. (3.10) to obtain the forecast error metrics for the respective method.

The extension to the FTSA case is straightforward. In one period, instead of comparing one univariate forecast and one univariate observation, an observed curve and a forecast curve are compared point-by-point. This procedure is repeated for the complete test set.

MSABC Comparing the forecasts with the observed curves requires a new measure. A point-by-point comparison was possible in the previous case only because the smoothed data were evaluated at the same points. Now, an integral-based measure is introduced to take into account that the curves were observed on irregularly spaced intervals and the forecasts exhibit a similar pattern. Hence, the new measure considers the inherent shape differences of the supply curve caused by bid size variations and length variations. The MSABC follows the idea that as the shapes of two curves are similar, the area between them is small, and, hence, the forecasting method provides a more precise prediction. For this purpose, let $y_t(x)$ be some observed curve and $\hat{y}_t(x)$ some forecast. Then, the MSABC for the evaluation interval $[\tau_0, \tau_L]$ is given by

$$MSABC = S^{-1} \sum_{h=1}^{S} \left(\int_{\tau_0}^{\tau_L} y_{\iota+h}(x)\, dx - \int_{\tau_0}^{\tau_L} \hat{y}_{\iota+h}(x)\, dx \right)^2, \tag{3.13}$$

where τ_0 denotes the lower limit and τ_L the upper limit.

The observed curve and the forecast may intersect, leading to a biased MSABC. To correct for intersections between observed curve and forecast, the MSABC incorporates the absolute area between intersection points if they are present.[9] Suppose there are $L-1$ intersection points denoted by τ_1 to τ_{L-1}. Then, let τ_l with $l = 0, 1, ..., L-1, L$ collect the upper and lower evaluation interval limits along with the intersection points. Then, the corrected MSABC is given by

$$MSABC = S^{-1} \sum_{h=1}^{S} \left(\sum_{l=0}^{L-1} \left| \int_{\tau_l}^{\tau_{l+1}} y_{\iota+h}(x)\, dx - \int_{\tau_l}^{\tau_{l+1}} \hat{y}_{\iota+h}(x)\, dx \right| \right)^2. \tag{3.14}$$

[9] The intersection points are calculated by utilizing *simple features* standards from the `sf` package in `R`. Simple features standardize the storage and representing of two-dimensional shapes.

With no intersections, Eq. (3.13) is a special case of Eq. (3.14).

The integral of the observed curve or forecast, i.e., the area under the curve, is either approximated by a stepwise connection of two points (Step) or by linear interpolation (Linear). The two approaches follow different ideas. The Step approach considers the inherent step function shape of the supply curves. However, a forecasting method might produce a curve not covering the whole interval from τ_0 to τ_L, i.e., the curve is truncated at the end, beginning, or both. In this case, the Step approach only uses the available points within a curve, whereas the Linear approach interpolates the truncated section by the closest available value of the curve. The forecasts are also obtained under the Ahead and Rolling principle.

3.4 Market and data

3.4.1 German balancing market

The power grid requires a constant balance between demand and supply achieved under the load-frequency control concept. The concept of load-frequency control allows to ensure physical delivery of electricity by correcting load deviations in case of under- or oversupply. But this is challenged by the increased penetration of renewable energy resources, which requires the maintainer's approach to be flexible, as argued by Alhelou et al. (2018) in their extensive review.

Van der Veen and Hakvoort (2016) define the balancing market in general as "*the institutional arrangement that establishes market-based balance management in an unbundled electricity market*". In Germany, the four TSOs – 50hertz, TenneT, Amperion, and Transnet BW – procure balancing power products in order to stabilize the grid, if there are deviations from the target frequency of 50 Hz, and are organized as *grid control cooperation* (GCC). As a member of the *European Network of Transmission System Operators for Electricity* (ENTSO-E), the GCC maintains the balancing market pursuant to Article 18(5) of Commission Regulation (EU) 2017/2195 of November 23, 2017, and the MfRRA.[10] The tendering is intended to be open, transparent, and free from discrimination according to guidelines of the Federal Cartel Office (Bundeskartellamt, BKartA), the national regulatory authority (Bundesnetzagentur, BNetzA), and the EU.

The TSOs organize a multi-unit pay-as-bid auction, and potential providers have to undergo pre-qualification covering the ability of applicants to provide the products in the required quality, e.g., the activation time, and to guarantee the information exchange with the responsible TSO. Currently, there are 57 pre-qualified providers. The auction follows a merit-order principle for a cost-friendly procurement accepting ascending price offers until the pre-announced demand is met.

There are three types of balancing power procured, which differ in their activation time:

[10]The MfRRA are guidelines for supplier of balancing power and are available in version 2.MfRRA of 2020/11/02 on the website of the GCC, `Link`.

- Frequency Containment Reserve (FCR, formerly: primary control reserve) with an activation time of 30s,
- Frequency Restoration Reserve with automatic activation (aFRR, formerly: secondary control reserve) with an activation time of 5 min, and
- Frequency Restoration Reserve with manual activation (mFRR, formerly tertiary control reserve) with an activation time of 15 min.[11]

After the announcement of the required capacity of balancing power by the TSO, each pre-qualified service provider can submit a bid that includes a capacity in Megawatt (MW), a *capacity price* [€/MW], and *energy price* [€/MWh]. The capacity price is paid for the provision of balancing power, and the energy price is paid for actually deployed power. The capacity is also referred to as the bid size.

As of December 1, 2007, the tender is made jointly by all four TSOs, and from July 27, 2011, until July 12, 2018, the auction was held weekly according to Decision Nr. BK6-10-098 of the BNetzA. The required capacity was quarterly announced. The aFRRs are distinguished by positive products (for excess of power consumption) for the participants to provide reserves or negative products (for lack of power consumption) for the participants to take power. They are labeled POS and NEG, respectively. Additionally, the aFRRs are procured for different periods. The abbreviations HT represents Monday to Friday from 08:00 to 20:00 and NT represents Monday to Friday 20:00 to 08:00 as well as Saturday, Sunday, and holidays 00:00 to 24:00. So, the providers can offer aFRRs for different categories. A category combines a product (POS, NEG) and a period (HT, NT). The bids are accepted according to the capacity price, and the activation follows the energy price, where the TSOs aim to achieve a cost-friendly ascending order for both cases.

Recently, there have been a few changes for this market. Since July 12, 2018, according to Decision Nr. BK6-15-158 of the BNetzA, the auction is held daily, and the periods are changed from HT and NT to 4h-blocks from 00:00 to 24:00 for all days of the week for POS and NEG. Since December 9, 2019, the required capacity is not quarterly announced but for each product individually. Furthermore, on November 2, 2020, an additional but closely related market under the name balancing energy market was introduced due to article 16 (5) EB-VO allowing the provider of an accepted bid to adjust the energy price after the initial auction.

The GCC is legally required to publish the anonymized data of the supply curves on their platform, regelleistung.net, according to the mentioned legal documents and §9 of the electricity grid access ordinance (StromNZV). Due to the publication obligation, the data is available, but the changes of the publication format make data processing a necessity. This project contributes large data sets with a detailed distinction between the supply and

[11] This terminology follows the European guidelines and was implemented by the GCC in June 2018 in the MfRRA, Sec. 4.9.

demand side for the literature on balancing power markets and the prediction of functional time series. For future analysis and a clean procedure, it uses the `tidyverse` (Wickham et al., 2019), a widespread collection of packages for data science sharing an underlying design philosophy, grammar, and data structures.

3.4.2 Descriptives

The complete data set of this project consists of 3.01 million observations and covers the period from June 2011 (2011/06/27) to November 2020 (2020/11/02) from the supply side. One observation is the (anonymous) provider's bid for one category (either POS or NEG and one period). The TSOs represent the demand side by their announced quantity for each category and the procured supply.

This project contributes three data sets. The primary data set, `balancing_market`, collects the bids of pre-qualified providers for each category of the aFRR. The second data set, `demand_ann` covers the quarterly demand of aFRR and mFRR that the TSOs announced on their platform. The last data set covers the procured supply of aFRR in `demand_obs`. All data sets, the codebook, and interactive 3D models are available online at the OSF.[12]

As mentioned before, the analysis is limited to the aFRR due to its relevance. In particular, the present study uses the capacity price of the POS HT aFRR of the weekly auctions from 2011/06/27 to 2018/07/09 ($n = 33,506$ for 368 weekly auctions). Fig. 3.3 illustrates the development of the capacity price over time for a selection of 20 observed supply curves that are temporally equally spaced and unprocessed. The price spikes that are common phenomena of the electricity market are also present in the balancing power market. There seems to be a downward trend starting in 2014, which is also found by Hinderks and Wagner (2019) in their analysis of the German day-ahead electricity market. The commonality illustrates the close relationship and dependencies between the two markets.

One supply curve shows the capacity price along the axis of accumulated bid sizes (0 MW to 2000 MW) up to the announced demand in Fig. 3.3. The last element of a supply curve is the point from which the TSOs do not accept any further bids and thus truncate the supply curves. The remaining (not accepted) bids are not publicly available. Hence, market participants cannot include all offers in their bidding strategy. The 20 examples in Fig. 3.3 show that the supply curves exhibit not only between-curve variation but also within-curve variation. The solid black line along the time axis (Jun 11' to Jul 18') represents the capacity prices of each week's first bid, and the black points mark the first offer of the 20 examples. The graph visually confirms the price spikes that lead to the between-curve variation, as the first capacity price defines the shape of the curve. Each weekly curve is colored by a blue to red scheme with 21 shades in increasing order, as shown at the bottom of the figure. The color represents the bid size. The non-constant but

[12] The online depository is online reachable by `Link` or Online (2022h).

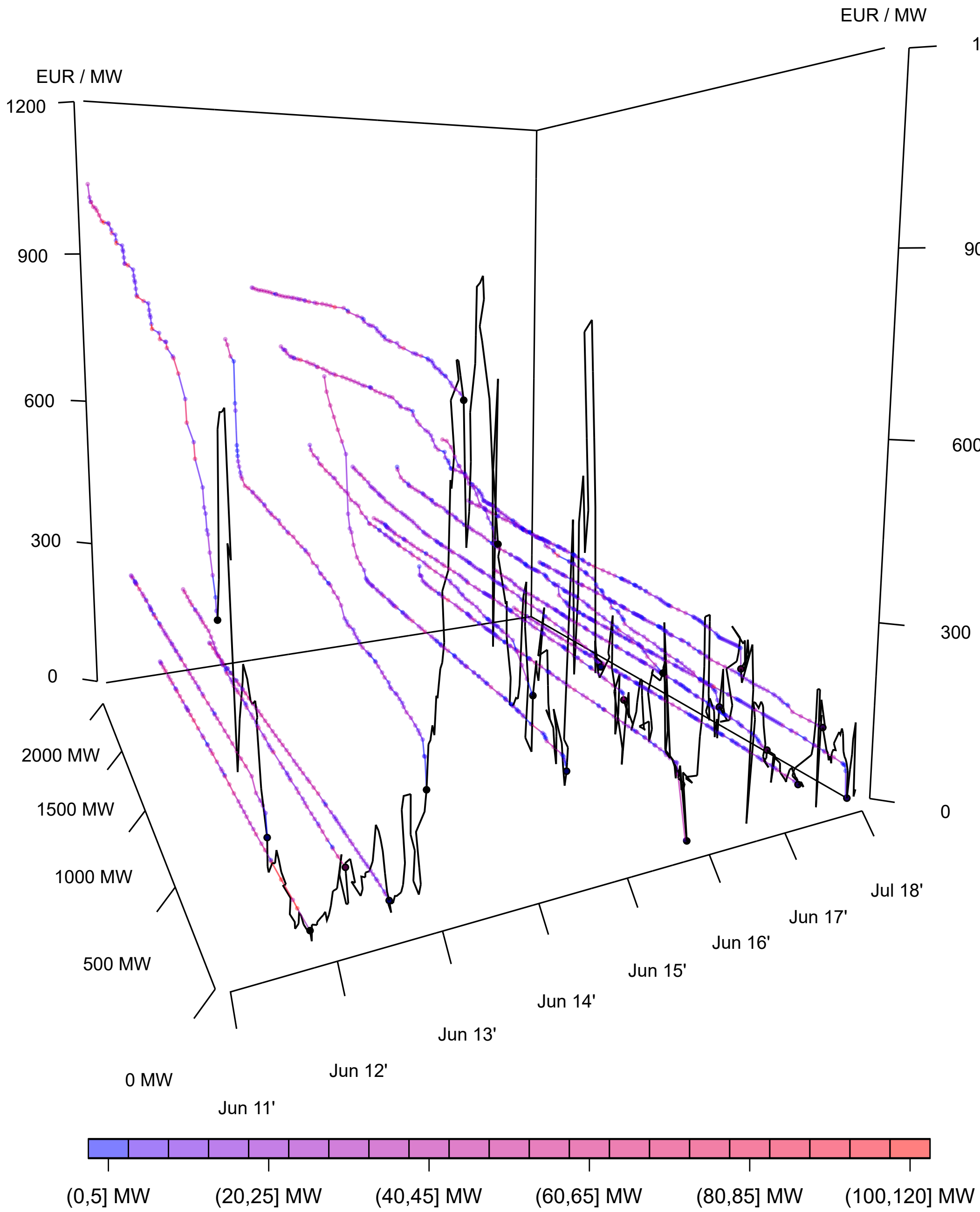


Figure 3.3. Supply curves for the German balancing market from 2011/06/27 to 2018/07/09 for weekly aFRR POS HT.

Note: The sample consists of only $n = 20$ temporally equidistant curves for visibility reasons. Each curve maps the cumulative MW supply to its respective capacity price, where the prices are in increasing order. The solid black line marks the price of all first bids. The black dots mark the first price for each of the 20 sample curves. The coloring scheme from blue to red with 21 shades represents bid size intervals and illustrates the unequal MW bid sizes across the periods along the prices axis by the non-homogeneous color gradient. The average capacity price for the weekly auction is 293.20 €/MW, the average offered capacity is 22.32 MW, and the average number of bids, i.e., the sequence length, is 91.05.

changing color gradient of each curve highlights the within-curve variations. The observed jumps within the curves leading to the different slopes imply that suppliers demand very different prices and vary the bid sizes. This substantial variation in the offered capacity is confirmed by the empirical cumulative distribution function of the bid sizes in Fig. A.1. It shows that capacities smaller than 20 MW barely cover 50%, and bids sizes of up to 50 MW cover roughly 94% of all observed capacity sizes in the weekly auction.

Moreover, the supply curves differ in their sequence length, i.e., the weekly number of bids, for two reasons. For one, given the wide range of dependency on electricity, the need for regulation lets the demand vary quarterly or daily, partly determining the length of a sequence. Secondly, the auction format does not regulate the price levels in their increment sizes or the upper limit of capacity offers, which leads to the slope variation. Fig. A.3 shows the observed sequence lengths and provides visual evidence for these problems.

Table 3.2 summarizes the weekly auctions of the POS HT aFRR. The average capacity price is 293.20 €/MW with a standard deviation of 220.65 €/MW. The smallest capacity price is 0€/MW and the highest 1531 €/MW. The average offered capacity is 22.32 MW with a standard deviation of 17.73 MW. The demand ranged from 1869 MW to 2500 MW. The TSOs accepted 91.05 bids on average with a standard deviation of 23.77. The weekly auctions achieved the announced demand by accepting at most 148 bids. Considering all procurements that were hold, the smallest number of accepted bids is 41. The smallest capacity was 2 MW and the largest 300 MW. Table A.1 in the appendix provides more descriptive statistics for the other product categories.

Table 3.2. Descriptive statistics of POS HT.

Variable	N	Mean	Std. Dev.	Min	$q_{0.25}$	$q_{0.75}$	Max
Capacity price	33506	293.20	220.65	0	128.23	367	1531
Offered capacity	33506	22.32	17.73	2	8.00	35	300
Demand	368	2032.64	99.87	1869	1973.00	2091	2500
Number of bids	368	91.05	23.77	41	70.00	112	148

Note: The table provides descriptive statistics for the weekly auctions of POS HT. The column $q_{0.25}$ shows the first quartile and the column $q_{0.75}$ shows the third quartile. The capacity price is given in €/MW, the offered capacity (bid size) and the demand is given in MW.

Overall, the curves vary between each other and within themselves. These characteristics force market participants to incorporate volume risks, undisclosed competitor behavior, and extreme price movements in the development of their bidding strategy and decision-making processes. Under these circumstances, reliable forecasts allow to efficiently allocate resources and maintain effective balancing management, guaranteeing a higher degree of security-of-supply. However, a consequence of these variations is that supply curves are observed under irregular spacing, making data processing necessary before applying some forecasting methods.

3.4.3 Data processing

This section provides details on the data processing that some benchmark methods in Section 3.3 require. To this end, let formally a provider's bid i in period t be given by

$$(q_{it}, p^c_{it}, p^e_{it})\,, \tag{3.15}$$

where q_{it} represents the offered capacity in MW, p^c_{it} the capacity price in [€/ MW] and p^e_{it} the energy price in [€/MWh] for $t = 1, \ldots, T$ and $i = 1, \ldots, n_t$. The analysis covers the capacity price of the aFRR in POS HT of the weekly auctions from 2011/06/27 to 2018/07/09. The sample is split into training ($\iota = 338$) and test set ($S = 30$), where the latter covers the supply curves from 2017/12/18 to 2018/07/09 and serves as the basis for computing the evaluation metrics in the Results section. In general, the prices are min-max normalized to the range from 0 to 1. The minimum and maximum are found within the training set.

Univariate case In the univariate case, a measure to summarize the supply curve is required. For this purpose, the *average weighted weekly capacity price* (AWWC price) for each week is computed. That is

$$\bar{p}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} \frac{q_{it}}{\sum_i q_{it}} \cdot p^c_{it}. \tag{3.16}$$

The AWWC price is illustrated in Fig. A.2, and the shape is similar to the solid black line in Fig. 3.3.

Input for the RNAA Typically, forecasting techniques based on neural networks utilize the Window approach for seq2seq prediction, which interprets the time series as a sequence along the natural time axis. However, for the RNAA, the sequences follow the bid order of the providers within one period. This fact highlights an essential difference in understanding sequences compared to other approaches. In a first step, this can be illustrated using the simplest sequence form for the univariate case. Here, the simple sequence, $\mathbf{y}_t$, on which the RNAA operates, is defined as

$$\mathbf{y}_t = (start,\ \bar{p}_t,\ end)\,, \tag{3.17}$$

where *start* and *end* are numbers illustrating the start and end of the sequence, respectively. So, for each period, a sequence of 3 steps is created. In contrast, the Window approach would generate sequences collecting multiple prices/periods, in general. Eq. (3.17) originates from the field of NLP, where a network aims to solve the task of translating sentences in two different languages. To this end, the probability of matching words of two dictionaries is optimized. However, for the network to understand the start and end of the sentences, both dictionaries are extended by two unique characters, words, or symbols placed at the start and end of each sentence.

The RNAA receives two inputs and one target output. By the imposed autoregressive input data structure, there is one input, $\mathbf{y}_t$, and one target sequence, $\mathbf{y}_{t+1}$, with each consisting of three steps. The implementation of teacher forcing in an efficient manner requires that the target sequence, $\mathbf{y}_{t+1}$, is split into two shorter sequences – the second input $\mathbf{y}^{in}_{t+1}$ and the target output $\mathbf{y}^{out}_{t+1}$. First, the target sequence is duplicated, and then one target sequence is reduced by the *end* entry to generate the second input, $\mathbf{y}^{in}_{t+1}$. After that, the *start* entry in the second target sequence is omitted to finalize the target output, $\mathbf{y}^{out}_{t+1}$. This format is common in NN projects and generated for each period t. So, the RNAA receives during the training process the first input $\mathbf{y}_t$, the second input $\mathbf{y}^{in}_{t+1}$, and the target output $\mathbf{y}^{out}_{t+1}$ defined by

$$\begin{aligned} \mathbf{y}_t = (start,\ \bar{p}_t,\ end)\,, \qquad \mathbf{y}^{in}_{t+1} &= (start,\ \bar{p}_{t+1})\,, \\ \mathbf{y}^{out}_{t+1} &= (\bar{p}_{t+1},\ end)\,. \end{aligned} \tag{3.18}$$

In the functional case, the AWWC prices $\bar{p}_t$ and $\bar{p}_{t+1}$ are replaced by the smoothed supply curve introduced in the following. In the supply curve case, the AWWC prices are replaced by the non-smoothed non-truncated supply curves consisting of the accumulated bid sizes and the capacity price.

Functional case The supply curves suffer from unequally spaced observations within each curve, as discussed before, and are truncated by the announced demand. The latter problem can easily be shown when looking at the announced demand and the supply over time in Fig. A.5. The demand varies over time due to multiple factors, so the supply curves are observed up to different MW levels. The maximum demand for the present case is 2500 MW, while the smallest is 1869 MW. Consequently, all supply curves are only jointly observed in the range from 0 MW to 1869 MW. Hence, the following functional data analysis needs to be limited to observations up to 1869 MW. For performance reasons, the prices were logarithmized (e.g. Klæboe et al., 2015) after an offset of +1 was added to each price controlling for prices of 0. The steps of the MWs were normalized to the interval $[0, 1]$ by representing each bids' capacity as a fraction of 1869 MW. These curves were smoothed as common in functional data analysis (Ramsay et al., 2009) to account for the discrete observations.

Each curve was smoothed by a cubic penalized spline (P-spline; Eilers and Marx, 1996) and constrained to be monotone increasing following Meyer (2012). P-splines are widely used, see Eilers et al. (2015), and are less computationally challenging as they use a discrete penalty matrix. The method relies on equally spaced knots and a B-spline basis while the penalty term, λ, governs the smoothness of the curve. For each curve, the optimal λ is found by a leave-one-out cross-validation (CV) scheme. Meyer (2012) proposed a shape-constrained P-spline estimator obtained by a weighted projection of the data onto a polyhedral convex cone. The parameters were chosen by the smallest average generalized cross-validation (GCV) value. Each smoothed curve was evaluated at 100 equally spaced points for further analysis. As in the univariate case, the sequences are extended by a *start* and *end* character. Hence, the approaches are supplied with smoothed and jointly truncated supply curves.

In the univariate case, it is common practice to check for stationarity of the time series before estimating a model. In the context of functional time series Horváth et al. (2014) propose a test with the null hypothesis that the functional time series is strictly stationary. For the smoothed data, the null hypothesis cannot be rejected with a p-value of 0.277 (Kokoszka and Shang, 2017).

Window approach The RNAA competes with simple NN models that rely on the Window approach for an additional evaluation step. In order to use a toolbox that implements the Window approach, the supply curves must be transformed to a seemingly two-dimensional time series. Hence, the supply curves are stacked in time, so the x-axis represents the time dimension, where the intervals between two dates are subdivided by the bid numbers. Fig. A.7 shows these stacked time series for the capacity price and the cumulative MW in the top and bottom panel, respectively. So, one window (blue) captures a supply curve as input and a second window (red) captures the consecutive target supply curve. The bid number is the position of a capacity within a supply curve determined by the increasing order of the capacity prices. Since each bid, see Eq. (3.15), consists of a capacity and a capacity price, both values share a common bid number, so the Window approach can correctly collect them in two separate series, as shown in Fig. A.7. Then, this technique moves both windows simultaneously ahead along the time axis by the identical number of steps for the next input and target pair.

Supply curve case In the final case, the RNAA receives the supply curves as a two-column matrix. For example, the input has the form

$$\mathbf{y}_t = \begin{pmatrix} start & start \\ q_{1t} & p^c_{1t} \\ \vdots & \vdots \\ \sum_i^{n_t} q_{it} & p^c_{n_t t} \\ end & end \end{pmatrix}, \tag{3.19}$$

where the first column is the accumulated bid size, given by q_{1t} for the second step, and $q_{1t} + q_{2t}$ for the third step, and so on. Note that the notation in this paper is not adapted for this case. For technical reasons, in `TensorFlow`, all sequences have the maximum length, but a very small negative value (-1e-12) fills the sequences after the *end* character.

3.5 Keras Tuner - optimal hyperparameters

The performance of machine learning methods crucially depends on the hyperparameter choice. With the almost infinite number of configuration possibilities, the need for a systematic approach arises. Furthermore, the growing size of networks and the availability of extensive data sets result in long computation times and complex interactions between and within layers. These are almost untraceable, making it difficult to evaluate the exact impact of parameter changes and their point of effect, rendering a brute force approach in most cases unfeasible.

The RNAA architecture applied here depends on the following hyperparameters. Within each layer, the cells have *activation functions*, which govern the gates of LSTM or GRU cells or compute the output. One common example is the *logistic sigmoid* function,

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{3.20}$$

which is applied element-wise on vectors in the present case. In general, this functional specification limits the output to the range of $(0, 1)$. Further examples of similar activation functions are the *hyperbolic tangent* (tanh) or the *softmax* function, or the *rectified linear units* (ReLU) function. The choice of activation functions in Bahdanau et al. (2015) is kept as far as necessary in the RNAA. Experimental trials show that in some cases, different functions lead to better but in some others also to worse performance. The output activation function of the Decoder and the deep learning layers were changed to the *linear* function to account for the range of the supply curves.

After the gradient has been acquired by *back-propagation* (Rumelhart et al., 1986), there exists a multitude of methods for the network to learn, which is the process of adjusting the weights of a network to map inputs to outputs, such as the famous *stochastic gradient descent*. A number of hyperparameters define the behavior of these learning methods as well, such as the *learning rate*, i.e., the step size at each iteration, and method-specific parameters to adjust for learning rate scheduling the later fine-tuning of the weights in the optimization process.

Moreover, the model designer has to choose the number of units within a cell of each layer, often referred to as the *latent dimension* and closely related to the output dimension of layers, which can differ for each layer type or be normalized to a single parameter. In addition, *initializers*, i.e., the starting values of weight matrices, bias terms, and the initial states of cells, if present, and the number of samples fed to the network in one iteration step, the *batch size*, need to be specified. The *depth* of a neural network refers to the number of layers it incorporates, and one *epoch* is one training step, where the complete training data are passed forward and backward through the model. In the RNAA context, the depth counts the number of deep learning layers.

The RNAA is programmed as part of the `Keras`[13] environment, which comes with a number of benefits. During the network building process, the architect has access to pre-defined functions, such as standardized fitting and evaluation functions and tools for visualizations and reporting. Furthermore, as `Keras` is one of the most used top-level frameworks in machine learning applications, new features are added steadily and are easily integrated into the RNAA. With the recent advances and updates of `TensorFlow`, a new package was introduced, which allows the systematic optimization of the hyperparameters.

[13]For more details, consider the official website `Link` or Online (2022f).

The `Keras Tuner`[14] works with a *HyperModel*, i.e., the set of all potential combinations of selected hyperparameters, and lets the architect specify a search for an optimal model, that is, the model with the smallest evaluation metrics, in a sophisticated way. The currently available search methods, labeled as *Tuner*s, include the *Random Search*, *Bayesian Optimization*, and the *Hyperband Search*, where the last extends the successive halving algorithm and is based on principled early-stopping to re-allocate available resources (Li et al., 2018).

The search results can be illustrated under mild restrictions in a three-dimensional space, as shown in Fig. 3.4, where the Tuner operated on the smoothed and jointly truncated curves (functional case). The parameters that define the Tuner's HyperModel restrict the search space. First, the latent dimension was normalized to be the same in each layer: in the Encoder, the Decoder, the attention mechanism, and the deep learning layer(s).[15] Then, the number of units within each layer was limited to the natural range of 8 up to 32. The depth of the RNAA could vary between non-existent (zero) or at most amount to five additional deep layers. The Tuner could switch between two learning algorithms, `Adam` (Kingma and Ba, 2015) and stochastic gradient descent, with three learning rates of 0.01, 0.001, and 0.0001. Finally, the Tuner was allowed to allocate 20 epochs during its search. In the first step, it tries to train as many models as possible and then reduces the number of initially considered models by a constant factor of 2. Fig. 3.4 displays the best 25 model trials, but their training duration, the number of epochs, differ. Hence, the figure provides only tendencies of the effect of the configuration concerning the loss on the validation set and aims to indicate the potential and relevance of hyperparameter tuning. Given more evaluated models, the points generate a surface and, if trained for the same duration, illustrate the impact of changes. However, due to the random initialization of weights, states, and biases, the reported results of a Tuner might differ between trials and even platforms. Fig. 3.4 provides hints that comparably small (latent dimension) and moderate deep models represent a good choice for a good performance.

3.6 Results

The performance of the RNAA is evaluated in three steps. In the first step, the RNAA is compared to the ARMA-GARCH model class (univariate case). Additionally, the attention plot is introduced, which is the visualization technique that comes with the RNAA. The second step shows the results when the smoothed and jointly truncated supply curves form the basis for estimation and forecasting (functional case). In the last step, the RNAA operates directly on the supply curves with varying sequence lengths (supply curve case).

[14] The official website can be found at `Link` or Online (2022e).

[15] In the notation of Appendix A.1, that relates to $n_e = n_d = n_a = n_p$.

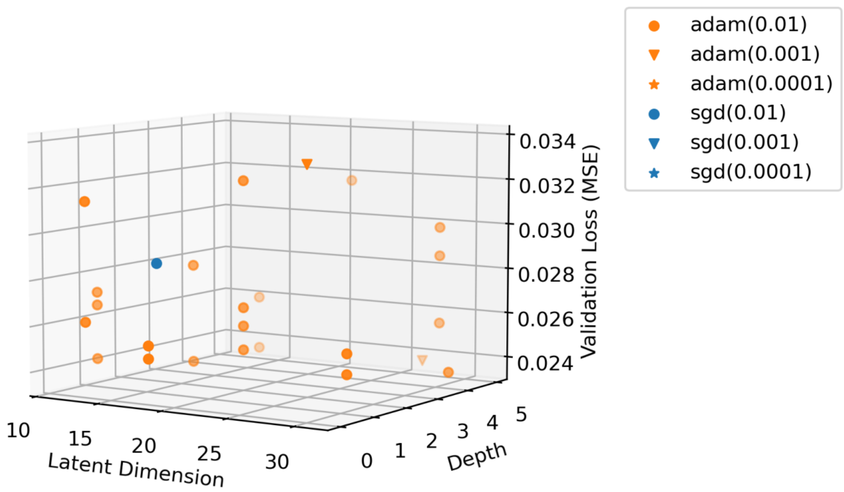


Figure 3.4. `Keras Tuner` - illustration of the 25 best trials in the functional case.

Note: The Tuner was the Hyperband search optimization on the smoothed and jointly truncated supply curves for the weekly auctions. The evaluation metric was the validation loss of the model as MSE. There are two learning methods, `Adam` (adam) and stochastic gradient descent (sgd), with three different learning rates, illustrated by symbols: dot (0.01), downward-facing triangle (0.001), and star (0.0001). The models are identified by their latent dimension and their depth but trained for different numbers of epochs according to the Tuner round.

3.6.1 Univariate case and the attention plot

In the univariate case, the RNAA uses a share of 0.237 of the training set for validation.[16] The latent dimension shared by Encoder, Decoder, attention mechanism, and Deep Learning is 28. The optimizer is `Adam` with a learning rate of 0.01. The Deep Learning consists of 5 additional deep learning layers. The model is trained for 30 epochs. The batch size is flexible and determined automatically during the training process. These hyperparameters were selected by a Hyperband search with the `Keras Tuner` in the restricted search space of Section 3.5. The tuner was also employed on different data formats. The *start* character is -1, and the *end* character is -2.

The estimation of the ARMA-GARCH model is based on the first differences, and the algorithm proposed by Hyndman and Khandakar (2008) provided the initial order of the ARMA(p, q) process. The subsequent grid-search for the order of the GARCH(r, s) with $r, s = 1, \ldots, 10$ yields an ARMA$(1, 1)$-GARCH$(1, 1)$ chosen by the *Bayesian information criterion* (BIC) and with appropriate properties. The differenced forecasts are retransformed to levels by adding their aggregated sum to the last observation of the training set.

Table 3.3 collects the evaluation metrics in the univariate case for the Naive, the ARMA-GARCH, and the RNAA.[17] It presents the MSE and the MAE under the two

[16] The number of observed curves is $n = 368$. The training set consists of 338, of which 80 form the validation set for the RNAA. The test set consists of 30 curves for all approaches.

[17] The benchmark models can be found in Table 3.1.

forecasting principles. The RNAA has a smaller MSE than Naive, i.e., repeating the last observed AWWC price under the Rolling and Ahead principle. This holds true for the MAE as well. The RNAA is also more precise than the ARMA-GARCH approach. This is even more striking due to the fact that the simple sequences from Eq. (3.17) force the RNAA into learning to predict the *end* character simultaneously while both other methods are operating on the univariate AWWC price time series. Moreover, the table reports only the deviations of predicted and observed AWWC prices for the RNAA. When the deviation between the predicted *end* characters and the true *end* characters is also included in the calculation, the RNAA's evaluation metrics become even smaller.

Table 3.3. Forecast accuracy evaluation (univariate case).

	MSE		MAE	
Method	Rolling	Ahead	Rolling	Ahead
Naive	0.00089	0.00864	0.02165	0.08474
ARMA-GARCH	0.00195	0.00609	0.03696	0.07007
RNAA	0.00085	0.00234	0.02132	0.03976

Note: The table summarizes the MSE and MAE between the true and predicted AWWC price under the Rolling and Ahead principle. The predictions are based on the Naive method, the ARMA(1, 1)-GARCH(1, 1), and the RNAA. The RNAA is trained with the AWWC prices in their simple sequence form, see Eq. (3.17). The other two operate on the AWWC prices directly.

After the RNAA has been sufficiently long trained, its attention mechanism and the attention weights assigned to the processed input can be visualized. This visualization is frequently labeled as *attention plot*, and in the translation context, it illustrates relations between a given input and its translation similar to a correlation heatmap. From an introductory perspective, this tool is best presented now in the univariate case. The attention plot is a two-dimensional plot that shows the attention weights of the attention mechanism averaged over the training set. It summarizes the average share, in other words, the assigned attention during training, of each step of the input sequence for every single step of the target sequence, so the row sum in the plot equals 1.

In Fig. 3.5, the attention plot for the RNAA in the univariate case is depicted. It shows the importance the attention mechanism assigns to each step of the input to predict a target step. Since the sequences are short, the numbers in the rectangles mark the size of the attention weights, and additionally, the coloring highlights more important steps in a darker shade. To control for outliers, the color range is truncated at the 99.5% quantile of all observed weight sizes, and larger weights than the quantile are grouped in the largest category. The displayed attention weights are each computed by averaging all weights for

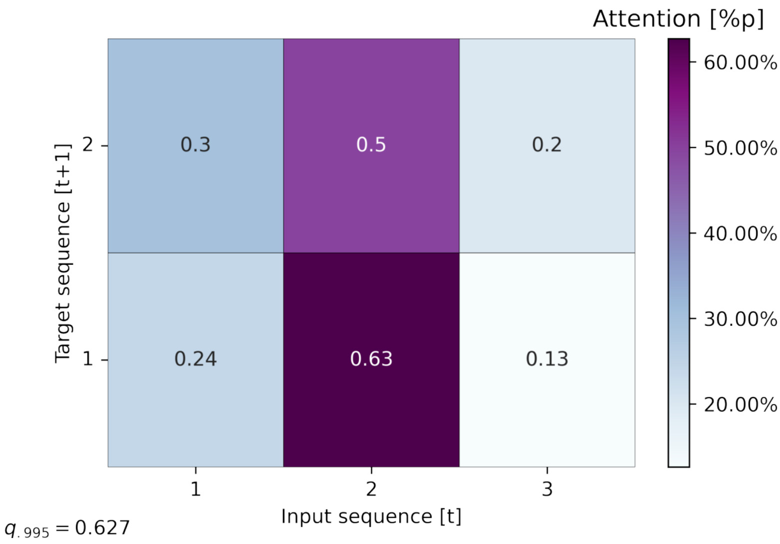


Figure 3.5. Attention plot (univariate case).

Note: Each sequence consists of the average weekly weighted capacity price and the *start* and *end*. The model was trained on the data from 2011/06/27 to 2017/12/11. The attention weights displayed in the attention plot are the average of all input and target training sequences. Additionally, the coloring illustrates from light to dark shade the amount of attention assigned in ascending order. Attention weights over the 99.5% quantile are grouped in the darkest category for visualization reasons.

one step pair of the target and input sequences in the training set.

It can be seen that the RNAA correctly allocates the most attention (0.63) to the second step of the input sequence, which represents the current AWWC price, $\bar{p}_t$ to predict the future AWWC price, $\bar{p}_{t+1}$, which is the first step of the target output sequence; see Eq. (3.18).

3.6.2 Functional case

In the functional case, the RNAA again uses a share of 0.237 of the training set for validation. The latent dimension shared by Encoder, Decoder, attention mechanism, and Deep Learning is 31. The optimizer is again `Adam` with a learning rate of 0.01. The Deep Learning consists of 4 additional deep learning layers. The model is again trained for 30 epochs. The batch size is flexible and determined automatically during the training process. These hyperparameters were selected by a Hyperband search as presented in Section 3.5. The *start* character is 0, and the *end* character is 1.

The simple NN benchmark models, namely Dense, CONV, and LSTM are also trained by the `Adam` optimizer with a learning rate of 0.01 to ensure comparability. The `TensorFlow` library comes equipped with an early-stopping functionality for the learning process. It stops the weight adjustment when it detects no favorable changes in the monitored metrics, e.g., when the decrease in a metric is below a pre-defined threshold. The simple NNs can learn for at most 20 epochs. However, the granted amount of epochs was never fully exploited since the learning was always stopped earlier.

The forecast error metrics for the functional case are summarized in Table 3.4. Obtaining the Naive forecast follows the previous section, but it consists of taking the complete past curves instead of the scalar values of past AWWC prices. The RNAA is more precise than the simple NN models, which have trouble performing better than the Naive forecast. The RNAA has lower metrics than the Bosq estimator, which shows better prediction than the Naive forecast. Only under the Rolling principle, the FTSA methodology performs better than the RNAA. Under the Ahead principle, the table is turned, and the RNAA shows more precise forecasts than the FTSA. Under the former principle, one potential driver for this result might be that the RNAA, despite the Rolling principle, feeds, for each step, its own prediction of the previous step in its computation procedure which causes an accumulation of errors. In particular, the teacher forcing requires a second input which is the predicted next step, and the inherent step-by-step processing of sequences of an RNN is leading to this collection of errors for each step of the target sequence.

Table 3.4. Forecast accuracy evaluation (functional case).

	MSE		MAE	
Method	Rolling	Ahead	Rolling	Ahead
Naive reptition				
Naive	0.00280	0.01236	0.02631	0.08350
Neural networks				
Dense	0.00304	0.13503	0.03227	0.31910
CONV	0.00497	0.16722	0.05191	0.37502
LSTM	0.01539	0.02570	0.11103	0.14653
Functional time series models				
Bosq	0.00248	0.01164	0.02721	0.09135
FTSA - Uni	0.00221	0.00983	0.02418	0.07965
FTSA - Multi	0.00228	0.00834	0.02106	0.07095
Proposed model				
RNAA	0.00247	0.00700	0.02619	0.06985

Note: The table summarizes the MSE and MAE between the smoothed curve and the predicted curve under the Rolling and Ahead principle. The predictions are based on the Naive method, the simple NNs (Dense, CONV, LSTM), the functional time series methods (Bosq, FTSA - Uni, FTSA - Multi), and the RNAA. The simple NNs are using a Window approach implementation. All predictions are based on the smoothed supply curves. Each smoothed curve was evaluated at the same points allowing the MSE and MAE to be calculated point-by-point.

Fig. 3.6 displays the attention plot of the RNAA in the functional case. It indicates that larger weights are around the beginning of the input sequences, implying that the starting price of the sequence determines the shape of the supply curve. Overall the deviations are small, but since the sequence length is around 100, the sum of impacts of one step of the input, which is the vertically collected sum of weights, on the complete target curve is worth mentioning. The attention plot does not reveal sizeable movements in the attention weights for most of the remaining sequence. Conceptually, the RNAA provides attention weights for each couple of input and target sequences. So for each couple, an attention plot is creatable, and it could be possible to identify more precise patterns in these individual plots. However, a few hundred plots might be challenging to interpret.

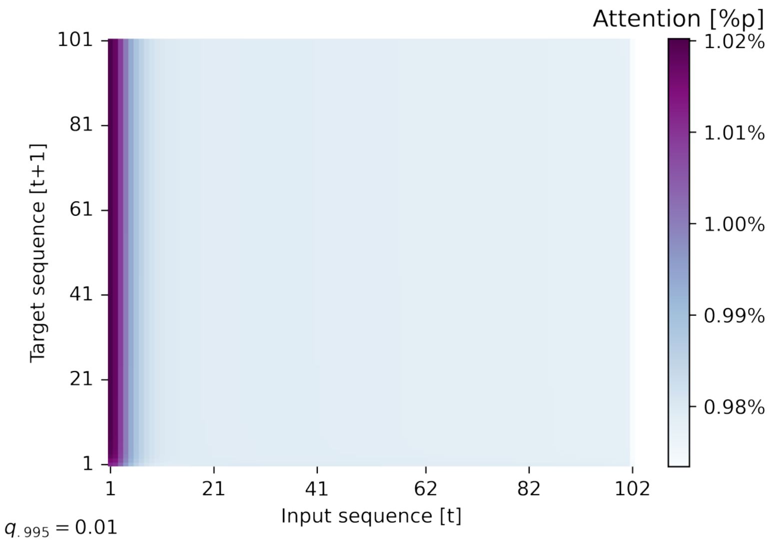


Figure 3.6. Attention plot (functional case).

Note: The sequences consist of the smoothed and jointly truncated supply curves evaluated at 100 equidistant points. The attention weights displayed in the attention plot are the average of all input and target training sequences. The model was trained on the data from 2011/06/27 to 2017/12/11. The coloring illustrated from light to dark shades the amount of attention assigned in ascending order. Attention weights over the 99.5% quantile are grouped in the darkest category for visualization reasons.

3.6.3 Supply curve case

In the supply curve case, the RNAA shows its full capabilities. In a preliminary step, Appendix A.3.1 shows how the RNAA outperforms simple NN implementations in almost all comparisons. In a next step, the RNAA competes with FTSA approaches. While the FTSA approaches require smoothed curves to be evaluated at the same points to get the same number per curve, the RNAA directly operates on the supply curves. Moreover, before smoothing, the curves were constrained to a common range. Thus, the FTSA methods are not able to predict beyond this range. In contrast, the RNAA is not restricted this way.

The comparison of FTSA and RNAA methods consists of two scenarios. The first scenario covers predicting the complete curves from start to end. It allows the procuring parties involved a more qualified decision as the forecast serves as a good cost estimate. In the second scenario, the focus shifts to the upper quartile of the curves. Here, the supplying parties involved might aim to place bids to maximize the revenue. To this end, the bids must be within the acceptance range, i.e., being part of the curve but in the more profitable last quarter. In both scenarios, the RNAA shows excellent performance.

The configuration of the RNAA is the result of extensive testing of various possible combinations. Compared to the other cases, the model is minimal. The training process uses a share of 0.237 of the training set for validation. The RNAA has a latent dimension of 16 and four additional dense layers to account for a deep learning structure. The optimizer `Adam` is responsible for training with a learning rate of 0.01. The model is trained for 25 epochs. The batch size is flexible and determined automatically during the training process. The *start* character is omitted, and the *end* character is -1e-12.

Table 3.5 summarizes the MSABC for this case under the Rolling and Ahead forecasting principle approximated by Step and Linear. At this point, it must be emphasized again that the RNAA uses the non-truncated supply curves. The FTSA methods (Bosq, FTSA - Uni, FTSA - Multi) use the truncated supply curves, i.e., $\tau_L = 1$ while the metrics for RNAA and the Naive method are computed up to the $\tau_L = \tau_t$, where $\tau_t \geq 1$ is the maximum demand as a share of 1869 MW in a week t. Thus, the RNAA is in a worse position from the start.

Despite that drawback, the RNAA shows the best performance by providing the smallest MSABC, as shown in the top panel in Table 3.5 for the total range. The MSABC is computed for all methods starting at $\tau_0 = 0$. Bosq shows no improvement compared to Naive but performs even worst. FTSA - Multi constantly outperforms the univariate approach FTSA - Uni. Compared to the best of the other approaches (FTSA - Multi), choosing the RNAA can reduce the MSABC by 11.9% (Rolling, Linear) and 13.9% (Ahead, Linear).

The picture is slightly different for the upper quartile. The MSABC is computed for all methods starting at the average of the weekly third quartiles, $\tau_0 = 0.7598$. Each week, the third quartile is the point (share of 1869 MW) where 75% of the bids are placed below. Given the announced demand, this calculation incorporates the bid sizes and the number of bids. Under the Rolling principle, the Naive method bests all other methods, except the FTSA - Multi in the Linear column. Here, both achieve a similar performance of 0.00006. The RNAA also achieves this performance if one evaluates its predictions up to $\tau_L = 1$, as done for the FTSA - Multi, instead of the actual supply curve end. Of course, this adjustment would also decrease the RNAA's MSABC for the total range; for more details, see Table A.3. Under the Ahead principle, the Naive method performs remarkably well, again. The only method that outperforms the Naive one is the RNAA. It can reduce the MSABC of the Naive method by 36.1% (Ahead, Linear).

The RNAA's attention mechanism produces attention weights based on the training set

Table 3.5. Forecast accuracy evaluation (supply curve case).

		Rolling		Ahead	
Range	Method	Step	Linear	Step	Linear
Total	Naive	0.00195	0.00192	0.00922	0.00915
	Bosq	0.00219	0.00207	0.01179	0.01117
	FTSA - Uni	0.00185	0.00171	0.00956	0.00904
	FTSA - Multi	0.00181	0.00168	0.00815	0.00770
	RNAA	0.00163	0.00148	0.00777	0.00663
Upper quartile	Naive	0.00012	0.00006	0.00053	0.00036
	Bosq	0.00020	0.00008	0.00077	0.00056
	FTSA - Uni	0.00018	0.00007	0.00063	0.00044
	FTSA - Multi	0.00016	0.00006	0.00057	0.00038
	RNAA	0.00019	0.00007	0.00027	0.00023

Note: The table summarizes the MSABC between the true curve and the predicted curve under the Rolling and Ahead principle. The FTSA methods (Bosq, FTSA - Uni, FTSA - Multi) are evaluated up to 1 [truncated supply curve]. The RNAA is evaluated up to the true end of the supply curve [non-truncated supply curve]. The column *Range* defines the starting point, i.e., the total range is from $\tau_0 = 0$ to $\tau_L \geq 1$, which depends on the methods. The upper quartile starts at the average upper quartile of 0.7598. The area under the curves is approximated by a stepwise connection of two points (Step) or by a linear interpolation (Linear).

as in the previous sections. Fig. 3.7 displays the corresponding attention plot. Contrary to the attention plot in the functional case, this plot shows that the RNAA considers the last third of the input sequences more relevant for the complete target sequence than the remaining parts. Moreover, the deviations are visibly higher than in the functional case, where the focus has been on the price dimension. In the supply curve case, the focus partly shifts to the capacity dimension since each step is a bid consisting of the capacity price and the capacity. Combined with the attention plot, this leads to an interesting insight: when using the supply curve's price dimension only and the forecast is restricted to the prices without considering the irregular spacing given by the bid sizes, valuable information might be lost.

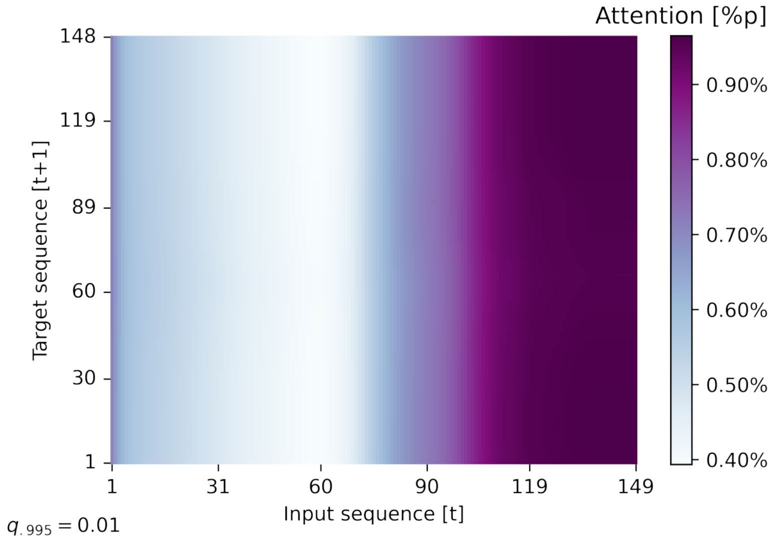


Figure 3.7. Attention plot (supply curve case).

Note: The sequences consist of the supply curve as they were observed. The attention weights displayed in the attention plot are the average of all input and target training sequences. The model was trained on the data from 2011/06/27 to 2017/12/11. The coloring illustrated from light to dark shades the amount of attention assigned in ascending order. Attention weights over the 99.5% quantile are grouped in the darkest category for visualization reasons.

3.7 Note on implementation

This section briefly discusses some remarks regarding the implementation of the RNAA in `Keras`. `Keras` is an API that makes model building straightforward and has a large, helpful, and active community along with extensive documentation and development guides. It runs on top of `TensorFlow` 2.0, an open-source library for differentiable programming developed by GoogleBrains, which has a focus on deep learning and provides scalability and cross-platform capabilities, such as execution on CPU, GPU, TPU, or even mobile devices. The routines are designed for `Python` 3.7, and care was taken to ensure that the results are reproducible by restricting random initialization and preventing multi-threading. The latter limits the computation power of devices significantly. But since the parameter number of the RNAA is small compared to, for example, image classification networks that come with the installation of `Keras`[18] and moderately deep, the RNAA is trainable on a personal device in reasonable times. On an i7-6700K Intel-based desktop, the search for an optimal model in the univariate case was conducted in 15 minutes and the initial epoch, in which the graph of the model is set up, took 7 seconds while each consecutive training epoch took less than 1 seconds. In the functional case, the Tuner took for the search approximately 1 hour, and the initial epoch was computed in 13 seconds and each additional epoch in 4 seconds. In the supply curve case, the epoch time of the training increased to 8 seconds. So, to summarize, the computation times are within a very appealing range, mostly due to the compactness of the RNAA, which is still scalable for better performance or more complex cases, while more potent hardware or cloud solutions keep the execution times short. The Decoder with attention mechanism was implemented so that it could be used in a `TensorFlow` 2.0 and easily added to other models. Moreover, this project provides classes for data preparation, visualization, and analysis. Given the fast pace of change in `Python`, the reproducibility and usability are limited to a similar environment to the one it was developed in, which is easily created by package management tools such as `Anaconda`. The RNAA uses the functional API of `Keras` for the multiple input functionality, and in the terminology of `TensorFlow`, the Attention-Decoder GRU, the joint `Python` implementation of GRU and attention mechanism, is a cell class developed in this project. An RNN layer loops through this cell for each step of a sequence.

3.8 Conclusion

This paper introduces the RNAA, a recurrent neural network with encoder-decoder architecture, an attention mechanism, and an imposed autoregressive input data structure. The RNAA learns to match an input sequence to a target sequence and their within-sequential dependencies while autonomously deciding which part of the encoded input sequence should be paid attention to under teacher forcing.

The performance of the proposed model is evaluated on data from the German balancing market. The RNAA improves in all benchmark cases the naive forecast approach. In the univariate case, where it is compared to an ARMA-GARCH model, it shows more precise predictions, and the introduced attention plot allows to gain insight into the relevance of sequence parts. In the functional case, the RNAA exhibits a convincing better performance which is extended to the supply curve case, where the observations are used without truncation or smoothing. In the last case, it easily outperforms simple NN approaches. Moreover, it allows discovering the relevance of multivariate bids for the prediction quality, and for the German balancing market, the middle and last segments of bids appear to be most relevant. To summarize, the RNAA is well suited for predicting supply curves.

The potential of the RNAA is not limited to this prediction task at hand but can extend and improve existing models. Staying in the world of electricity prices, one could imagine

[18] See the overview of famous networks at `Link` or Online (2022d).

building on the idea of Shah and Lisi (2020) and Ziel and Steinert (2018) by modeling the sales and purchase curves separately by an RNAA and also adding a mapping to the actual electricity price, which should deliver promising results.

The model can also easily be augmented for the inclusion of external predictors. When information for each step is available, it can be taken into account by including it as an additional column, as illustrated in the supply curve prediction section. Moreover, when external variables impact the complete supply curve but are only observed as a scalar in each period, the extension of the RNAA is straightforward. An additional attention mechanism can include, weight, and scale these covariates so that the RNAA has not only access to additional information but can also illustrate its relevance. As there is some consensus in the EPF that the inclusion of external predictors leads to better predictions, more precise prediction can be expected by these augmentations, and the presented results for the German balancing market show only a fraction of the potential of the RNAA. Given the wide range of application opportunities, this article hopefully contributes a flexible and intuitively appealing approach for future research.

Chapter 4

Round-number Effects in Bargaining: Bias vs. Focal Point[19]

It is a well-established fact that there is a variety of biases when it comes to numbers. There is ample evidence of left-digit bias (Busse et al., 2013; Englmaier et al., 2018; Lacetera et al., 2012), the role of prominent numbers in decision processes (Converse and Dennis, 2018), and the clustering of prices at round numbers in the real-estate market (Pope et al., 2015; Repetto and Solís, 2019) and the energy market (Shah and Lisi, 2020; Ziel and Steinert, 2016). There is a growing body of literature that recognizes the importance of precise ($1.67) and round numbers ($2.0) in decision-making. In particular, initial offers play an important role as anchors in the bargaining literature (Janiszewski and Uy, 2008; Loschelder et al., 2014). Mason et al. (2013) report making precise offers is a signal of being more informed compared to someone making a round offer. Hukkanen and Keloharju (2019) even advise not to initiate a bargaining process with a round number. Yan and Pena-Marin (2017) argue that round-number offers signal “completion” and “goal achievement” and are consequently linked to a higher acceptance propensity.

We study the role of round numbers in bargaining situations and whether their effect can be explained by preferences for round numbers (*round-number bias*) or by their role as a solution for a coordination problem (*focal point*). We hypothesize that faster decisions and higher acceptance frequencies result from a round-number bias, focal points, or both. To investigate this, we provide first empirical evidence of a negative correlation between the duration of a bargaining process and the usage of round price offers. For this finding, we exploit the vast data set of Backus et al. (2020), which covers over 11 million observations of eBay bargaining protocols with the so-called *Best Offer* option enabled. Our finding is robust and highly significant but does not allow for a causal interpretation of the role of round numbers.

For this purpose, we design an online experiment that incorporates a dynamic bargaining game. We record the decisions of participants to accept or reject a random offer that might be round or non-round. Based on our first treatment, we obtain a measure for

[19] This chapter is based on the joint work Lauf and Schlereth (2022).

the preference for round numbers as we eliminate any dependency on the decisions of others for the participant. In the second treatment, we keep the incentive structure, but additionally introduce a coordination problem where the effect of the participant's decisions now depends on the decisions of a second participant. Comparing the first and the second treatment sheds further light on the effect of round-number bias or the role of round numbers as focal points in the spirit of Schelling (1960). In our framework, this means that despite potentially having no round-number bias, participants might accept round offers because they believe their partners are more likely to accept round offers. We made sure that there is no reason for the round-number bias to be different across treatments, other than their possible role as tools for coordination.

Our experiment confirms the finding from the eBay data set that round offers facilitate faster acceptance. In addition, we find clear evidence of a round-number bias. Participants are more likely to accept round offers. Finally, we do find evidence for round numbers serving as focal points, but only under certain conditions. We find significant increases in acceptance frequencies of round numbers for female participants under coordination for less advantageous offers. No such effect can be seen for more advantageous offers. The offer's advantageousness is determined by its potential payoff for the participant. For male participants, the patterns look similar but are less pronounced and non-significant.

Our paper contributes to various strands of the literature related to negotiations, focal points, the salience of roundness, and round-number bias. In general, negotiations are instruments to find solutions for disagreements in various fields, such as trade, politics, and social life. We are interested in the potential role of round numbers in price negotiations. Here, round numbers might serve as focal points to reach agreements as introduced by Thomas Schelling. He argues that focal points constitute a solution not necessarily depending on logic but frequently on prominence or conspicuousness. When numbers are involved, outcomes show a strong tendency towards simplicity in the form of roundness (Schelling, 1960).

The first strand of literature investigates negotiations in an experimental setting. Recent evidence supports Schelling's argument that payoff-irrelevant but conspicuous labels for players' strategies, e.g., “Option A” or “Option B”, facilitate coordination in tacit bargaining, that is, situations in which communication is not possible. Coordination games can model such situations. For example, they may involve two players choosing a strategy from a common set and receiving a payoff only when they chose the same strategy (Mehta et al., 1994b). The literature found saliently labeled strategies can serve as focal points in one-shot coordination games (Bardsley et al., 2010; Crawford et al., 2008; Mehta et al., 1994a; Parravano and Poulsen, 2015). We embedded a form of a coordination game in our second treatment as participants must accept the same offer. Our design differs through its dynamic character and the fact that we place label and payoff in the same domain. The number 10 might not only be the number that lies equidistant between 9 and 11, but could also serve as a salient label.

Besides the salience of labels, other features and payoff asymmetry may influence coordination. For example, Isoni et al. (2013) introduce the *bargaining table* game and study spatial salience. They find that focal points increase efficiency in a tacit bargaining game, even when these cues induce unequal payoff divisions. Isoni et al. (2019) find that salient labeling increases coordination success, but they also document minor and major disruptive effects of payoff inequality and conflict of interest. In our experimental design, payoffs are symmetrical for now, but our design would allow us to easily incorporate asymmetric payoffs to study their impact on coordination further.

Empirical work has shown that the visual representation of numbers influences the outcome of economic decision-making. A prominent example is the literature on the left-digit bias. Lacetera et al. (2012) provide evidence for a left-digit bias of consumers in the wholesale used-car market. They report threshold effects at 10,000-mile odometer marks resulting in discrete price drops. Moreover, Busse et al. (2013) analyze retail data on used cars and arrive at the same conclusion. Finally, Englmaier et al. (2018) also report price discontinuities at salient mileage thresholds for the European market and extend the analysis to the age of the car. One possible explanation of these findings is an overestimation of the distance to the next round mileage when cognitive constraints result in only the leftmost digits being processed. For example, a car with 20,000km is perceived as far less valuable than one with 19,999km. We add to this strand of the literature by studying visually salient round numbers and their effect on decision-making in negotiations.

For this purpose, we conducted an empirical analysis of a recently published data set from a well-known platform - eBay. Besides auctions, eBay offers a platform for sequential bargaining, which is named “Best offer”. Backus et al. (2020) collected a data set of eBay transactions and listings. Their study focuses on comparing their results from the data with theoretical predictions from the bargaining literature. In Backus et al. (2019), they utilize the same data and find evidence for cheap-talk signaling in the use of round-number listing prices and offers. They document a trade-off. Round-number offers are on average lower but are more likely to sell. In our analysis, we focus on how round numbers affect the duration of negotiations.

There does, however, also seem to be a perceptional difference between round and non-round numbers, even in the absence of stressful situations. The psychological literature has long recognized the relevance of how numbers are presented, i.e., whether they are *round* or *prominent* (Converse and Dennis, 2018). Rosch (1975) finds that such round numbers serve as reference points in lab settings. Empirical findings for marathon runners (Allen et al., 2017), baseball, SATs, lab experiments (Pope and Simonsohn, 2011), and preventive health behavior (Wadhwa and Zhang, 2019) support this. Converse and Dennis (2018) provide evidence for round-number effects in financial market data and experiments. In five studies simulating real-world scenarios, such as buying coffee or selling a textbook, Yan and Pena-Marin (2017) discover that offers’ roundness increases the willingness to accept in experimental bargaining in line with their hypothesis that round numbers symbolize completion. Lin et al. (2020) analyze 2,000 classroom experiments where the simplest form of bargaining, an ultimatum game, was conducted and find clear spikes at offers that are multiples of 10. Our approach provides another perspective, covering cooperation and focal points. We contribute a novel design focusing on the decision to accept an offer without the influence of scenario-based stimuli or communication between subjects.

A synthesis of the strands of literature on focal points and round numbers is the study of Pope et al. (2015). They find evidence in support of round numbers serving as focal points in high-stake real estate negotiations. Still, they raise the question of to which extent a round-number bias or the role as a focal point is responsible for the relevance of round numbers in negotiations. Our paper addresses this question by providing a novel experimental framework to make a clearer distinction.

The remainder is structured as follows. Section 4.1 discusses our empirical analysis of the eBay data set and the motivation for our experimental design, which we present in Section 4.2. In Section 4.3, we present our experimental results. Section 4.4 concludes.

4.1 Empirical evidence from eBay

It is a common phenomenon that in real-world negotiations such as buying a used car or when participating in auctions and garage sales, the final prices surprisingly often tend to be round. We were interested in whether this tendency is also associated with other outcomes of the bargaining process, such as the acceptance frequencies and the swiftness of negotiations. We found such a relationship, for which we provide the empirical evidence in the following. For this purpose, we use the data from Backus et al. (2020), who made available millions of records of single-unit fixed-price listings from May 2012 to June 2013 on the US eBay site, where the “Best Offer” option was enabled.

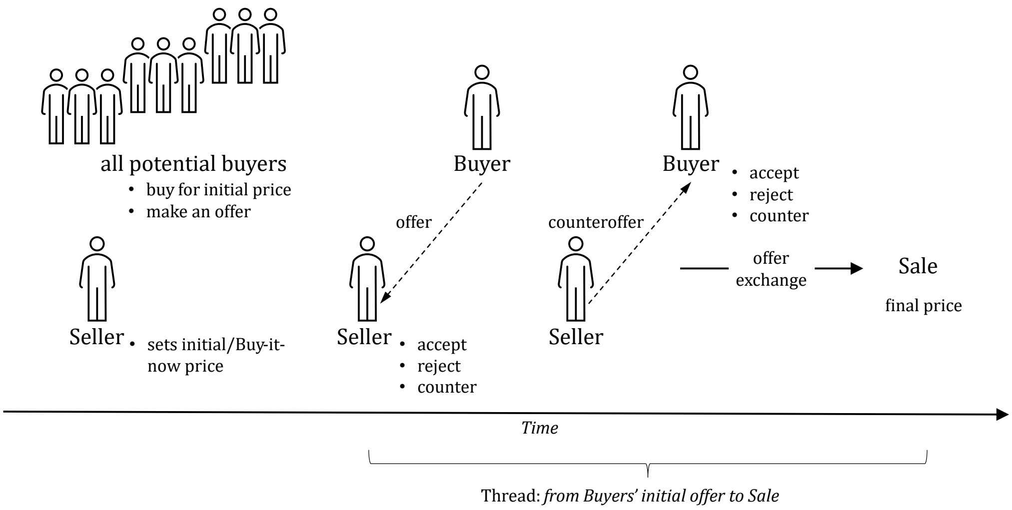


Figure 4.1. Illustration of the “Best Offer” option on eBay.

The procedure on eBay illustrated in Fig. 4.1 is as follows. First, a seller sets an initial price for the item to be sold, also called the “Buy-It-Now” price. Then, all potential buyers can instantly buy the item for this price or if the "Best Offer" option is enabled, send the seller an alternative price offer. Next, the seller can accept or reject this offer or make a counteroffer giving the buyer the same possibilities. Both can make at most three offers, and each is valid for 48 hours. Finally, the item is sold for the final price if both parties agree. Otherwise, the negotiations fail.

Backus et al. (2019) document that using round numbers is associated with lower prices for the seller as argued in the literature (Hukkanen and Keloharju, 2019; Janiszewski and Uy, 2008; Loschelder et al., 2014; Mason et al., 2013). However, Backus et al. (2019) argue that past research has ignored an important trade-off: a round price may come at the benefit of a higher likelihood of a sale. They use their own extensive data set and show for 10.5 million listings that a round initial price, in the form of multiples of \$100, increases the likelihood of a sale by 3%-points to 6%-points while a share of 20% of all listings is sold on average.[20]

We focus on a different aspect by analyzing the role of round numbers within successful negotiations instead of the signaling effect of initial prices. We find that the share of round numbers increases from 15.4% of the initial prices to 41.2% of the final prices. This might be driven by a round-number bias but could also be due to round numbers being used as focal points to accelerate a settlement within the bargaining process. If round final prices are indeed associated with faster settlements, this will provide evidence for either or a combination of both.

To this end, we use the two data sets provided by Backus et al. (2020). The first data set consists of *threads*, which are sequences of offers for one buyer-seller pair bargaining over one item as well as their responses, as shown in Fig. 4.1. Hence, one thread consists of multiple observations, but the last (most recent) observation covers the bargaining outcome and the price for which the bargaining parties settled. The second data set holds information on the items within the threads of the first data set, such as its condition and the category it falls within. We developed an algorithm that processed and merged both data sets. In particular, we collected the duration of each thread, its final price, and additional information on the sold item in a new data set. The duration is the time passed between the first and last observation within one thread, and we matched it with details on the corresponding item. Appendix B.1 provides more information about our algorithm. In total, we collected 11.1 million threads, and Table 4.1 summarizes the data.

For our empirical analysis of bargaining times, we introduce the following notation. Let i be the identifier of a successful thread, where the seller sold the item for the final price p_i. The thread's duration is captured by Δt_i and represents the time between the buyer's initial offer and the last observation in the thread, which either is the automatic or the manual acceptance of the seller or the manual acceptance of the buyer. The set of round prices is denoted by Υ, and $I_\Upsilon(p_i)$ is an indicator function that is equal to 1 if the

[20]The data set is restricted to *Collectibles* with an initial listing price between \$50 and \$550, where the round numbers are $z \in \{100, 200, 300, 400, 500\}$; more details in section IV.B.3 of Backus et al. (2019).

Table 4.1. Descriptive statistics of the eBay data.

	Mean	Median	SD	Min	Max
Duration (min)	1,049.07	136.60	7,948.14	0.00	802,791.77
Periods	1.51	1.00	0.92	1.00	9.00
Final price ($)	81.11	29.00	142.98	0.99	1,100.00
Round numbers (final price)	0.41	0.00	0.49	0.00	1.00
Initial price ($)	118.84	39.99	3,307.96	0.99	6,000,000.00
Round numbers (initial price)	0.15	0.00	0.36	0.00	1.00
Number of photos	3.26	2.00	2.99	0.00	12.00
Seller's feedback score (%)	99.67	99.86	2.10	0.00	100.00
Observations	11,090,279				

Note: The table summarizes the eBay data set of Backus et al. (2020) after applying our algorithm. The distribution of the items' conditions and categories can be found in Table B.1 and Table B.2.

final price is round, and 0 otherwise. Each thread relates to one item, and X_i collects any additional information on this item. We define *round numbers* as the 5-step intervals up to 50, followed by 10-step intervals up to 100 and extended by 50-step intervals up to 1000. In particular, let the set of round prices be given by

$$\Upsilon = \{5, 10, 15, ..., 45, 50, 60, 70, ..., 90, 100, 150, 200, ..., 950, 1000\}.$$

We estimate the model

$$\Delta t_i = \beta I_\Upsilon(p_i) + c + \boldsymbol{\gamma} X_i + u_i, \tag{4.1}$$

where i denotes an observed successful thread, Δt_i is the duration measured in minutes, p_i represents the observed final price of the item, c is a constant, and X_i collects the condition of the item (11 categories, baseline is "New", see Table B.1) and the meta category of the item (38 categories, baseline is"Collectible", see Table B.2). We only use threads with final prices of up to \$1100, covering 98.2% of the successful threads.

The resulting data set consists of 11.09 million threads with a successful sale, where the average initial listing price is \$118.84, and only a share of 15.4% of the initial prices is round. So, sellers started roughly every seventh thread with a round price. The parties settled on average at a final price of \$81.11, and intriguingly the share of round numbers increased to 41.2% of all final prices. The average duration of a thread is 1,049 min 4.5 s (SD: 7,948 min 8.34 s), but the median is 136 min and 36 s.

Table 4.2 shows the results of OLS regressions both without controls and with controlling for the item's condition and category. The effect of round numbers, captured in Eq. (4.1) by β, is reported in the row *Round numbers*. In all specifications, we find a significant negative relationship between the duration of a thread and the dummy indicating that the item was sold for a round price. In particular, on average, threads ending with a round final price were 53 min shorter than threads without a round final price (see our preferred specification in Column (2)). In the appendix, Table B.3 lists the results broken down for all conditions and categories.

Table 4.2. Regression results. Round final prices and the duration of successful eBay sales.

	(1) Duration	(2) Duration	(3) Periods	(4) Periods
Round numbers	-24.82***	-53.02***	-0.17***	-0.19***
	(4.91)	(5.21)	(0.00054)	(0.00064)
Constant	1059.31***	1165.92***	1.58***	1.59***
	(3.02)	(9.80)	(0.00038)	(0.0011)
Condition dummies	-	Yes	-	Yes
Category dummies	-	Yes	-	Yes
N	11,090,279	8,144,375	11,090,279	8,144,375

Note: The table reports OLS results for the two dependent variables, *Duration* and *Periods*.Duration denotes the time between the first observation and the last observation of a thread in minutes. Periods denotes the number of offers made between seller and buyer. The table reports the coefficient of the round number dummy as *Round numbers*. There are 11 condition dummies for the item, where the baseline is "New". The meta category of the item has 38 categories and is considered with a corresponding number of dummies, where the baseline is "Collectible". Missing observations are due to incomplete recordings of condition or category. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

We assess the robustness of our findings by different checks. First, we use another measure for duration and replace Δt_i by the number of periods of the bargaining process, i.e., how often each party made an offer. We observe an average number of periods of 1.48 in the data. We confirm our previous result because when the final price is round, the number of periods is reduced by 0.19 on average, controlling for the category and condition of the item (Table 4.2, Column (4)). Moreover, we applied a placebo test by shifting each element of $\varUpsilon$ by 1 ahead, i.e., $\varUpsilon + 1$. We find that the effect is insignificant in the duration case ($\hat{\beta} = -7.84(12.7)$, standard error in parenthesis) and the effect on the number of periods becomes very small ($\hat{\beta} = -0.005(0.0016)$), yet remains significant. We additionally checked the influence of available covariates in the data set, such as initial listing price, number of item's photos, and the seller's feedback score in Eq. (4.1), which did not change the results.

4.2 Experimental design and implementation

We illustrated the relevance of round numbers in a real-world bargaining setting, with the intriguing finding that the share of round numbers increases for successful negotiations, and that negotiations that end with a round-number price are shorter. With our experiment, we want to answer the question if, and to what extent, the acceptance of round numbers is driven by individual behavioral biases or by round numbers serving as focal points in negotiations. An experiment trying to answer this question must possess some essential characteristics. We summarize the desirable properties in the following:

Properties of the experimental design:

1. **One-player and two-player**: The design needs to be suitable for a one-player setup (where individual behavioral biases might kick in) and a two-player setup (where, additionally, focal points may play a role).

2. **One change at a time**: The extension to the two-player case must be possible with only one change at a time.

3. **No communication**: The channel of making offers that could serve as signals needs to be closed. Bargaining typically involves making offers to others, receiving offers, and evaluating counter offers by another (human) party. Strategic considerations might induce round-number effects already when making offers. Allowing participants to freely exchange offers would make an analysis of acceptance decisions rather difficult.

4. **Abstract environment**: Context-specific restrictions on the offers need to be eliminated since the bargain's item or the environment presented in the study might determine a particular set of reasonable offers.

5. **Offer size**: Round offers should not be more financially attractive than non-round offers.

6. **Upside of rejection**: To avoid that subjects simply accept every offer, there must be some value in rejecting a given offer (i.e., try to get a better offer than the current one).

7. **Downside of rejection**: At the same time, rejecting a given offer must be costly, so subjects cannot wait infinitely long (i.e., waiting incurs the risk of ending up worse than the current offer).

8. **Same number of decisions**: The design must allow eliciting the same number of decisions from each participant for the homogenization of the lengths. The reason is that negotiations generally end when both parties agree or break down when the continuation seems unattractive, leading to an inherent variation in the negotiations' lengths.

Hence, our experiment is necessarily somewhat abstract, but we argue that we have found a way to incorporate the most relevant characteristics of bargaining. Therefore, and in favor of clarity, we limit the following section, 4.2.1, to an outline of the procedure and postpone the detailed explanation of our design choices to the subsequent section, 4.2.2.

4.2.1 Experimental design

The basic idea of our design is as follows: We randomly present each participant an offer from a pre-defined offer set. Then, participants can either accept or reject that offer. Accepting an offer might result in a payoff corresponding to the offer's size.

We have two treatments, **Single** and **Partner**. In Single, each participant completes 10 periods. Each period consists of two steps, *Preview* and *Decision*. For Period 1, Fig. 4.2 shows the two screens that visualize the two steps. In the first step, Preview, participants see the offer set of the current period from which they will get a randomly drawn offer; see panel (a) of Fig. 4.2. Each offer is equally likely to be drawn within one period. In the second step, Decision, the participants receive an offer from the set displayed in the previous step. Then, they decide whether to accept or reject it. Additionally, they can see the offer set of the next period; see panel (b) of Fig. 4.2. After each period, the three largest offers are removed from the offer set. The next figure, Fig. 4.3, illustrates how in Period 2, the first line, Line 1, is removed from the offer set. In Period 3, the next line, Line 2, is removed (*shrinking cake* design). This design was explained to participants in detail prior to the start of the experiment.

A participant's payoff is the sum of the participation fee of \$1.50 and a bonus. For the bonus calculation, one period is randomly selected after the participants have completed the study. We call the selected period and all subsequent periods *payoff-relevant*. With this, the bonus equals the first accepted offer in a payoff-relevant period multiplied by the conversion rate, \$0.001. We added a hint about the dependency on the selected period on the Decision page; see panel (b) of Fig. 4.2 just above the buttons. At the end of the study, participants can view their decisions and the selected period.

In Partner, everything remains the same except for the bonus calculation. We match each participant with another participant whom we call the participant's partner. Both see the same sequence of offers across the periods. As before, a randomly selected period specifies the participant's and partner's payoff-relevant periods. However, in Partner, the bonus is the first offer that both players individually accept multiplied by the conversion rate. In other words, we introduce a coordination game in each period. If no offer in a payoff-relevant period is accepted by both players, they each receive a zero payoff. The hint on the Decision page is adjusted so that it additionally includes the dependency on the partner's decision.[21]

In both treatments, participants made their decisions individually and privately. We did not allow any communication. In Partner, participants have no information about their partner's identity or decisions. We made sure that participants understood the shrinking cake design and the role of payoff-relevant periods by a treatment-specific check-

[21] The hint can be found in Fig. 4.2 and was: Single: *Remember: Whether you get the offer also depends on the period the computer selects.*; Partner: *Remember: Whether you get the offer also depends on the period the computer selects* ***and the decision of your partner***. (Italic and bold font is added here for readability.)

(a) Preview

(b) Decision

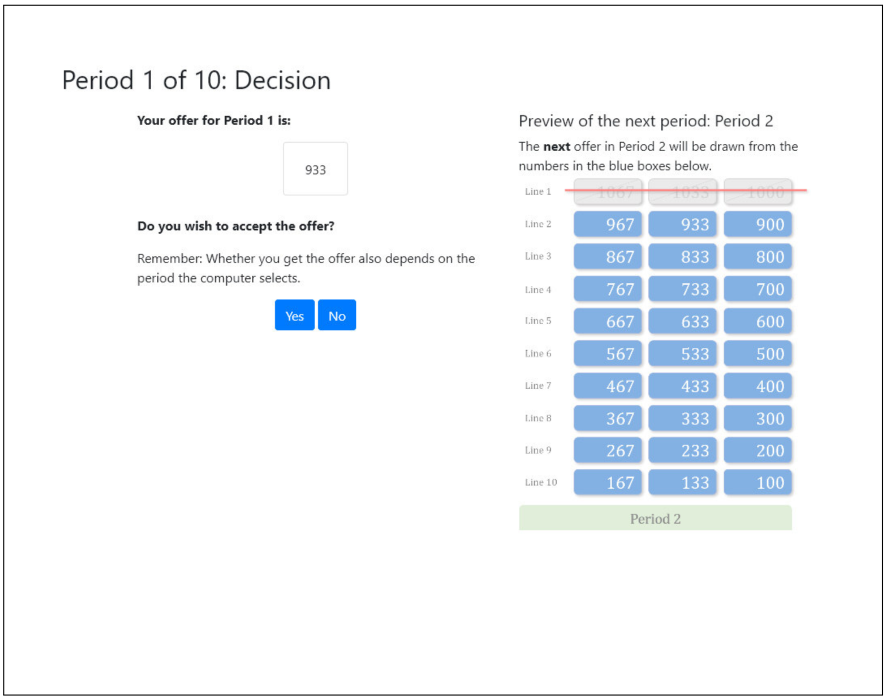


Figure 4.2. The two steps of Period 1.

Note: The top panel (a) is the step Preview which shows participants the offer set of the current period. The bottom panel (b) displays the step Decision where participants can accept or reject the randomly drawn offer in the black rectangle. The right side displays the preview of the next period. A red line marks the removed numbers, and the gray background highlights the removal. The subsequent periods have the same layout, except that additional lines are removed; see Appendices B.3.2 and B.3.3. Only three buttons (“Get your offer”,“Yes”, “No”) are clickable.

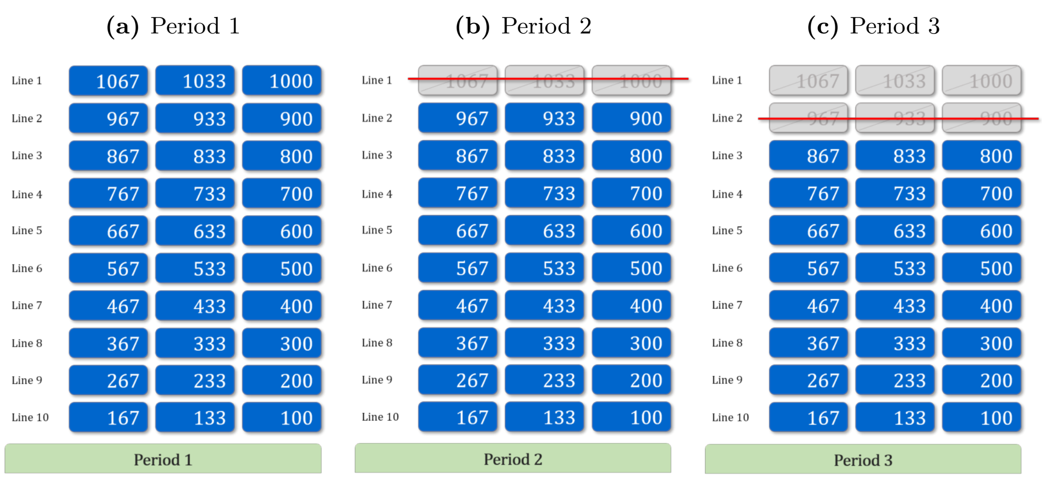


Figure 4.3. Offer set for Period 1, Period 2 and Period 3.

Note: All offers in blue boxes could be drawn with equal probability by the experimental software in the respective period. The green box at the bottom displays the current period. The removal of the numbered lines is illustrated by the red line crossing out 3 numbers and is highlighted by the gray boxes.

up questionnaire before the actual study. In order to control for an equal distribution of the random offers across treatments and participants, we formed groups of 4 participants who saw the same sequence of offers and assigned 2 participants to each treatment. We achieved randomization by sorting people into groups by their time of arrival.

4.2.2 Discussion of the required properties

The first property is easily fulfilled by our design as the participants in Single might exhibit a round-number bias, and Partner introduces coordination with a partner, where round numbers might serve as focal points. The extension to the two-player case is achieved by a single change of the bonus calculation, whereby everything else remains the same. This satisfies property 2.

Our design focuses on the decision to accept or reject an offer. Therefore, it intentionally does not allow communication between participants, such as freely making offers or exchanging messages. This way, we obtain property 3. On the one hand, the design prevents participants from sending their partner a signal by using numbers to transmit their intentions or hints about their future behavior. On the other hand, it enables a cleaner comparison between Single and Partner because, in the former treatment, there is only a single participant to elicit the individual round-number bias.

Although we acknowledge that it might be interesting and possibly relevant in observational data, we need to abstract from an open bargaining approach, including communication, to distinguish between round-number bias and round numbers as focal points while keeping the differences between both treatments minimal. Thus, we decided to make offers exogenous. This might, of course, result in us underestimating the role of round numbers as focal points as we omit their usage in the between-participants communication and focus on the decision-making. Nevertheless, making the offers exogenous comes with the additional benefit of having the identical offer sequences in Single and Partner.

All offers in our experiment will be drawn from the set of numbers in Fig. 4.3. We chose this set of numbers and its visual representation to have round and non-round numbers blend in naturally with each other. Hence, we obtain property 4. In addition, the intervals between the numbers are evenly spaced, and no number has decimal places. Moreover, contrary to other studies, we avoid any association of the offer set with specific situations, such as buying used cars, investing in stocks, or selling a house, since it might affect the perception of round numbers. For each round offer, there are at least two higher non-round offers. In particular, the average round offer is always smaller than the average non-round offer, so, e.g., in Period 1, the average round offer is 550, and the average non-round offer is 600. We will show in Section 4.3.1 that a large set of (standard) utility functions predicts lower acceptance frequencies for round offers (property 5).

Our bonus calculation allows us to obtain the remaining properties. Briefly summarizing our calculation, a randomly selected period determines the payoff-relevant periods, but the participants only learn which periods are payoff-relevant after having made all ten decisions. So, if a participant simply accepts every offer, an undesirable offer might be the first offer within the payoff-relevant periods and become the bonus. Hence, rejecting an offer might be advantageous, and accepting all offers is in general not the optimal strategy (property 6). However, rejecting an offer might be costly because the shrinking cake design causes the three highest offers to be removed after each period, reducing the chance of getting a better offer than the current one. Thus, there is also a potential downside to waiting for a better offer (property 7).

Since the period in which the decisions become payoff-relevant will be disclosed at the end of the study only, the participants should not consider previous periods at any decision. Should they have accepted an offer in a previous period that turns out to have been payoff-relevant, further acceptances have no impact on the bonus payment. So there is no downside in accepting further offers in future periods. However, as they can never be certain if a previous acceptance will indeed have been payoff-relevant, the decision in the current period could determine the bonus. The same is true, should they have rejected all previous payoff-relevant offers. Again, the decision in the current period could determine the bonus. Therefore, each decision is incentivized, regardless of previous decisions. Hence, participants should always behave as if the current period will be selected as the beginning of payoff-relevancy and only compare the current offer to possible future ones. The hint towards the payoff-relevancy on the Decision page reminded the participants of the bonus calculation. Additionally, due to the post-study disclosure of the selected period, accepting an offer does not lead to the end of the study. Still, all participants must complete all ten periods, which yields the same number of observations from each participant for each treatment and offer sequence (property 8).

4.2.3 Implementation

The experiment was programmed in `oTree` (Chen et al., 2016). The study was conducted on Amazon Mechanical Turk (MTurk henceforth) in December 2020 using a sample of MTurk experienced US residents. In total, 924 participants (382 women) took part in the experiment, earning $1.90 on average with an average completion time of approximately 7 minutes. After the experiment, participants had to fill out a short post-experiment questionnaire. Table 4.3 summarizes descriptive statistics of the sample.[22] The study was preregistered at the AEA RCT Registry under the ID AEARCTR-0006823.

Table 4.3. Descriptive statistics of the sample.

		Treatments		
	Total	Single	Partner	p-value
	(N=924)	(N=462)	(N=462)	
Gender				
Female	382 (41.3%)	198 (42.9%)	184 (39.8%)	0.350
Male	542 (58.7%)	264 (57.1%)	278 (60.2%)	
Age (years)				
Mean (SD)	37.9 (10.9)	38.0 (10.6)	37.7 (11.1)	0.540
Education				
Less than a high school degree	4 (0.4%)	3 (0.6%)	1 (0.2%)	0.692
High School Diploma	76 (8.2%)	37 (8.0%)	39 (8.4%)	
Vocational Training	7 (0.8%)	2 (0.4%)	5 (1.1%)	
Some College	82 (8.9%)	36 (7.8%)	46 (10.0%)	
Associate's degree	59 (6.4%)	29 (6.3%)	30 (6.5%)	
Bachelor's degree	501 (54.2%)	262 (56.7%)	239 (51.7%)	
Master's degree	177 (19.2%)	83 (18.0%)	94 (20.3%)	
Professional degree	14 (1.5%)	8 (1.7%)	6 (1.3%)	
Doctoral degree	4 (0.4%)	2 (0.4%)	2 (0.4%)	

Note: The table summarizes the characteristics of the subject pool. The column *Total* shows the number of observations for each category of gender, age and education. The columns *Single* and *Partner* report the distribution across treatments. The column *p-value* reports the p-value of tests between the two treatment groups. In particular, for Gender and Education, the χ^2 test was applied, and for Age, the Wilcoxon-Mann Whitney test.

4.3 Experimental results

In our analysis of the eBay data set (see Section 4.1), we showed that round final prices are associated with faster decisions when the negotiations ended with the sale of an item. Considering this finding, we start the discussion of our experimental results with a look at

[22]The announcement via which the participants were invited to the study can be found in Appendix B.3.1 and the instructions are summarized in Appendices B.3.2 and B.3.3.

the time participants needed to make their decision within our experiment. To this end, we define the *decision time* as the time needed to complete a period in seconds.

In a first step, we consider all decisions, acceptances and rejections alike, and Table 4.4 shows the average decision times for each treatment and the two offer types. Comparing the top and bottom row of the table shows that decisions were made quicker for round offers. When the observations are pooled across treatments, we find significantly quicker decisions when a round offer was made (t-test: 9.07s vs. 10.18s; $p = 0.0004$). The difference in decision times between the two offer types is 1.11s. When we control for the treatments, we find that participants in Partner, who received a round offer, decide significantly quicker (t-test: 9.07s vs. 10.62s; $p = 0.0013$) while the difference in Single yields a p-value of 0.1006 (t-test: 9.08s vs. 9.73s).

Table 4.4. Decision times.

		Treatment	
Offer type	Total	Single	Partner
Round	9.07	9.08	9.07
NonRound	10.18	9.73	10.62

Note: Average decision times are reported in seconds.

In a next step, we consider acceptances and rejections separately. When the observations of acceptances are pooled across treatments, we find significantly quicker acceptances when a round offer was made (t-test: 9.81s vs. 11.05s; $p = 0.0054$). In addition, the difference in decision times between offer types for acceptances is slightly larger (1.24s) than in the previous case (1.11s). When we control for the treatments, we find that participants in Partner accept round offers significantly quicker (t-test: 9.75s vs. 11.52s; $p = 0.0112$) while the difference in Single (t-test: 9.88s vs. 10.55s; $p = 0.2123$) is not significant. For the rejections, we observe a similar picture with a difference in decision times of only 0.73s but with shorter decision times in general. The corresponding tables and details can be found in Tables B.4 and B.5 in Appendix B.2.1.

Generally, there appears to be a difference in decision-making when round numbers are involved. Round numbers seem to trigger faster decisions. Since our offer set for each period is designed so that there are two larger non-round offers for each round offer, and the size of the offer is directly related to the potential payoff, round offers are on average less favorable than non-round ones. Hence, the decision times of rejections being smaller for round offers comes with no surprise, albeit the difference is very small and just significant. Intriguingly, round offers were also accepted quicker, leading to a more sizeable decrease in decision times between round and non-round offers than the rejections. We now turn to the likelihood that an offer was accepted and the role of round offers. For this purpose, we calculate the *acceptance frequency* as the fraction of participants that accepted a given offer. Alternatively, in more practical terms, this is the number of participants who clicked on "Yes" in the Decision step in panel (b) of Fig. 4.2 relative to all participants.

4.3.1 Acceptance frequencies

In a preliminary step, we show that within our experimental framework a large set of (standard) utility functions (such as risk aversion, risk neutrality, or quantal response) predict lower acceptance probabilities for round offers and hence would not be consistent with an experimental finding of higher acceptance frequencies for round offers. We will show that this holds before the start of the experiment by showing that it holds before any period.

To this end, let the ten periods be denoted by $t = 1, 2, \ldots, 10$. As outlined in Section 4.2.1, the top line of the offer set is removed after each period. Thus, the offer set in period t is a matrix $X_t = (x_{i,j})$, where $i \in \{1, 2, ..., m_t\}$, $m_t = 11 - t$ and $j \in \{1, 2, 3\}$. By construction, for all i, we have (a) $x_{i,1} > x_{i,2} > x_{i,3}$, (b) $x_{i,1}$ and $x_{i,2}$ are non-round numbers, and (c) $x_{i,3}$ is a round number.

For some offer $x_{i,j}$ from X_t, let $p_t(x_{i,j})$ denote the probability that $x_{i,j}$ is accepted in period t, given some utility function $u(x_{i,j})$, where the outside option, i.e., the expected utility of rejecting and waiting for some future period, is denoted by μ_t. From the perspective of the beginning of period t (i.e., before an offer has been randomly drawn), the probability that, given a round offer is made, it will be accepted is given by

$$P_t^R = \frac{\sum_{i=1}^{m_t} p_t(x_{i,3})}{m_t}. \tag{4.2}$$

Analogously, the probability that, given a non-round offer is made, it will be accepted is given by

$$P_t^{NR} = \frac{\sum_{i=1}^{m_t} \left(p_t(x_{i,1}) + p_t(x_{i,2})\right)}{2m_t}. \tag{4.3}$$

Proposition. *The probability that round offers are accepted is lower than the probability that non-round offers are accepted (i.e., $P_t^R \leq P_t^{NR}$) if $p_t(x_{i,j})$ is weakly increasing in $x_{i,j}$. In particular, this is the case if the agent has standard risk-neutral preferences, standard risk-averse preferences, or follows quantal-response behavior. In the case of quantal-response behavior, the above inequality holds strictly.*

Proof. Substituting Eq. (4.2) and Eq. (4.3) in $P_t^R \leq P_t^{NR}$ yields

$$\sum_{i=1}^{m_t} p_t(x_{i,3}) \leq \underbrace{\frac{m_t}{2m_t}}_{=\frac{1}{2}} \sum_{i=1}^{m_t} \left(p_t(x_{i,1}) + p_t(x_{i,2})\right). \tag{4.4}$$

Inequality (4.4) is satisfied if

$$p_t(x_{i,3}) \leq \frac{1}{2}\left(p_t(x_{i,1}) + p_t(x_{i,2})\right) \tag{4.5}$$

is satisfied for all i. The monotonicity of $p_t(\cdot)$ and $x_{i,1} > x_{i,2} > x_{i,3}$ imply that this is the case. □

$p_t(x_{i,j})$ is weakly increasing in $x_{i,j}$ if the decision-maker has a utility function $u(x_{i,j})$ that is weakly increasing in $x_{i,j}$ (which is, for example, the case under risk neutrality or risk aversion): a rational decision-maker would accept any offer $x_{i,j}$ where $u(x_{i,j})$ is weakly greater than the outside option μ_t. Thus, $p_t(x_{i,j}) = 1$ whenever $u(x_{i,j}) \geq \mu_t$, and $p_t(x_{i,j}) = 0$ otherwise.

For a decision-maker following quantal-response behavior, the probability to accept an offer is given by

$$p_t(x_{i,j}) = \frac{1}{1 + e^{\lambda(\mu_t - u(x_{i,j}))}}, \tag{4.6}$$

again resulting in monotonicity of $p_t(x_{i,j})$ in $x_{i,j}$. In this case, $p_t(x_{i,j})$ is strictly increasing in $x_{i,j}$ whenever $u(x_{i,j})$ is strictly increasing in $x_{i,j}$.

Having thus shown that $P_t^R \leq P_t^{NR}$ holds for every period t, it also holds in general.

4.3.2 Round-number effects

We now turn our attention to the experimental results as they relate to our research question. We are interested in the presence of round-number effects and possible differences between individual and cooperative decision-making.

Our analysis starts by estimating a simple linear probability model with the offer acceptance as the dependent variable. *Offer acceptance* is a binary variable equal to 1 if an offer was accepted or 0 otherwise. In a first specification, we regress acceptance on a dummy variable for the treatment and an indicator for round numbers. We also add the interaction of these two variables to allow for differences of round-number effects between treatments. The results of this regression are presented in Column (1) of Table 4.5.

In light of our theoretical predictions, we would expect round numbers to have lower acceptance frequencies, should there be no round-number bias. In Column (1), we do indeed find a negative sign for the round-number dummy, but the effect is not significant. This insignificance provides a first hint at the presence of round-number effects. We find a slight but significant increase in acceptance frequencies in Partner. The interaction term represents the additional effect of round offers in Partner. The negative sign implies that participants are less likely to accept round offers, but it is not significant.

For a more in-depth analysis, we control for the size of an offer. It seems quite reasonable to assume that, ceteris paribus, higher offers are more likely to be accepted. In our experimental set-up, however, the size of an offer has to be evaluated in relation to the current period. An offer of 233 might not be attractive in period $t = 1$, but much more attractive in period $t = 8$. We, therefore, need to control for the relative size of the offers in our analysis. Hence, we introduce the *offer share*.

The *offer share* is defined as $s_t = \frac{x_t}{x_{\max,t}}$, where x_t is the offer in period t and $x_{\max,t}$ is the largest possible offer in period t. Thus, for a given period, the offer share measures the relative size of a given offer compared to the largest possible offer. Given our parameters, it follows that $s_1 \in \left[\frac{100}{1067}, 1\right]$ in the first period.

Fig. 4.4 shows the acceptance frequencies for each observed offer share (gray dots)

Table 4.5. OLS Regression. Dependent variable: Offer acceptance.

	(1)	(2)	(3)	(4)
Sample:	Total	Total	Female	Male
(Intercept)	0.620 ***	0.230 ***	0.225 ***	0.235 ***
	(0.012)	(0.023)	(0.035)	(0.029)
Offer Share		0.596 ***	0.580 ***	0.608 ***
		(0.025)	(0.038)	(0.032)
Treatment: Partner	0.030 *	0.030 *	0.065 **	0.006
	(0.018)	(0.018)	(0.027)	(0.024)
Round Offer	-0.021	0.045 ***	0.056 ***	0.036 **
	(0.014)	(0.014)	(0.020)	(0.018)
Treatment: Partner x Round Offer	-0.010	-0.010	-0.016	-0.005
	(0.021)	(0.020)	(0.030)	(0.026)
N	9240	9240	3820	5420
R2	0.001	0.099	0.098	0.100

Note: The offer acceptance is a binary variable equal to 1 if the participant accepted an offer and 0 otherwise. Standard errors in parentheses are heteroskedasticity robust and clustered on the individual level. The results in columns (1) and (2) are based on the total sample. The results in columns (3) and (4) are based on the female and male sub-sample, respectively. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

and the evaluated fit (curve) of a Nadaraya–Watson kernel estimation where we pool the observations from both treatments. Not surprisingly, the curve illustrates that higher offer shares are more likely to be accepted. This allows us to investigate how round offers affect acceptance frequencies in more detail.

Consider Column (2) of Table 4.5. We add the *offer share* to our linear probability model. Unsurprisingly it is significant and drastically increases R^2. The coefficient on the round-number dummy now turns positive and is significant at the 1%-level. The other coefficients remain qualitatively and quantitatively unchanged. Thus, on average, ceteris paribus, round offers were about 4.5%p more likely to be accepted.

By separating the data by our two treatments, we can observe differences in how these round-number effects influence acceptance decisions. We split the data into four distinct categories: by treatment (Single, Partner) and by offer type (Round, NonRound). For each category, we estimate a Nadaraya–Watson kernel regression. Fig. 4.5 visualizes these estimates for each treatment (Single: gray, Partner: black) for round offers (solid line) and non-round offers (dashed line).

We observe an intriguing pattern. To illustrate the pattern more clearly, we present separate figures for each treatment. The left frame of Fig. 4.5 shows the Single treatment. For smaller offer shares, we observe no differences between round and non-round numbers. For higher offer shares, the solid line representing round numbers lies above the dashed

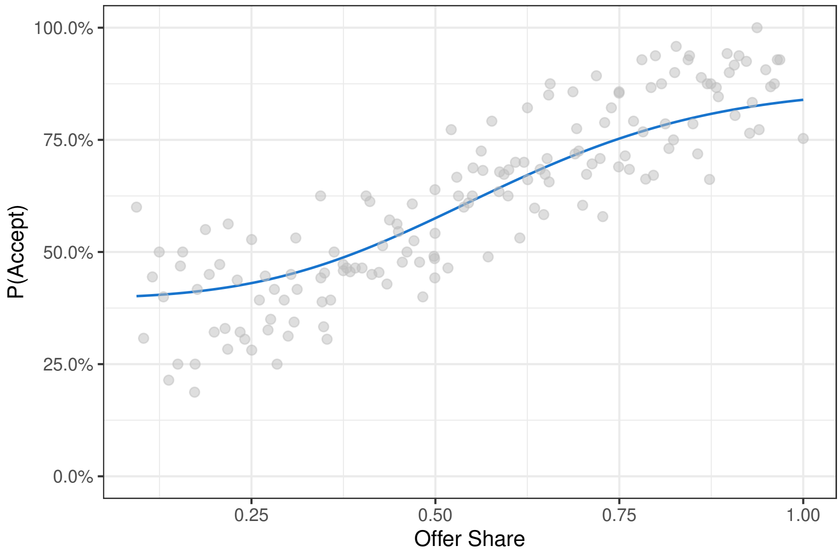


Figure 4.4. Acceptance frequencies for the pooled sample.

Note: The curve is based on a Nadaraya–Watson kernel estimator using a normal kernel with a bandwidth of 0.4. The gray dots represent the frequency of acceptance for a given offer share.

line representing non-round numbers. This implies that for higher offers, round numbers are more likely to be accepted by our subjects. Conversely, this does not seem to be the case for relatively low offers. In the right frame of Fig. 4.5, in Partner, the pattern is reversed. The solid line lies above the dashed line only for smaller offer shares, while no major differences can be seen for larger offer shares. Thus, if anything, there is a higher frequency of accepting lower round numbers. Nevertheless, in both treatments, the solid line (which represents round offers) is above the dashed line, which indicates that round offers are more likely to be accepted in general. We will refer to the differences between the acceptance frequencies for round and non-round numbers as *round-number effect* for now, as we will later disentangle whether this effect is driven by bias or coordination.

Over the last decades, studies in behavioral economics have shown that women and men exhibit different behavior in several economic domains. For example, it is often found that men are less risk averse than women, less charitable, and more competitive (see e.g., Croson and Gneezy, 2009; Niederle, 2016). It might be possible that there are also differences between men and women when it comes to bargaining and round-number effects.

There are more men than women in our sample (58.7% vs. 41.3%), but we do not find that one sub-sample received higher offers or more round offers or that men and women were unequally distributed into treatments..[23]

Consider now Columns (3) and (4) in Table 4.5. For both sub-samples, there is a positive round-number effect. The effect is stronger for the female sub-sample at 5.6%p

[23]The Mann-Whitney test on the offer size between the two sub-samples yields a p-value of 0.1832 and a p-value of 0.2728 for the same test on offer share. The treatments are independently distributed across gender sub-samples (χ^2, $p = 0.3497$). The number of round offers is independent of the gender (χ^2, $p = 0.7796$).

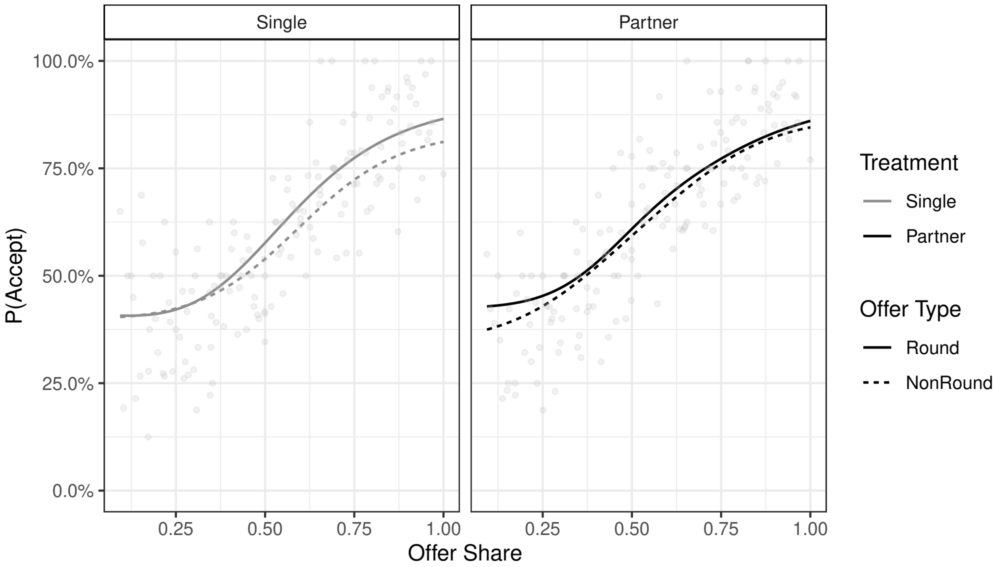


Figure 4.5. Acceptance frequencies for the treatment and offer type sub-samples.

Note: The curves are based on a Nadaraya–Watson kernel estimator using a normal kernel with a bandwidth of 0.4. The gray dots represent the frequencies to accept for a given offer share. The gray lines represent the Single treatment in the left frame, and the black lines correspond to the Partner treatment in the right frame. The offer types are illustrated with solid lines for round offers and dashed ones for non-round offers.

with significance at the 1%-level. For the male sub-sample, the coefficient is 3.6%p, and is significant at the 5%-level. Women also react differently to the treatment than men, showing an increase in acceptance frequencies in Partner of 6.5%p at the 5%-level. The interaction term is insignificant in both sub-samples.

In a next step, we repeat the exercise from before to analyze the effect of non-round and round offers in both treatments by splitting each gender sub-sample into four distinct categories and estimating a curve for each of these. A visual analysis confirms that there are indeed behavioral differences between men and women.

The curves in Fig. 4.6 visualize for each category the acceptance frequencies, where the left panel presents the female sub-sample and the right one the male sub-sample. Differences with respect to gender are clearly visible here. For the female sub-sample the pattern we observed in Fig. 4.5 for the pooled sample is even more pronounced. Again, the solid line in Single lies above the dashed line for higher offer shares, while in Partner, it lies above the dashed line only for lower offer shares. In general, for the female sub-sample, round offers are more likely to be accepted than non-round offers, as indicated by the fact that the solid line is almost always above the dashed line in each treatment. In the male sub-sample, the curves are very close to each other and almost overlapping. If anything, there is a tendency for the solid line to be above the dashed line, implying that round offers are more likely to be accepted. Analogous to the analysis without controlling for gender, men and women are more likely to accept round offers than non-round offers in both treatments. However, the intriguing patterns of the round-number effects in Fig. 4.5

appear to be mainly driven by the female sub-sample. Therefore, we focus on the female sub-sample for further analysis.

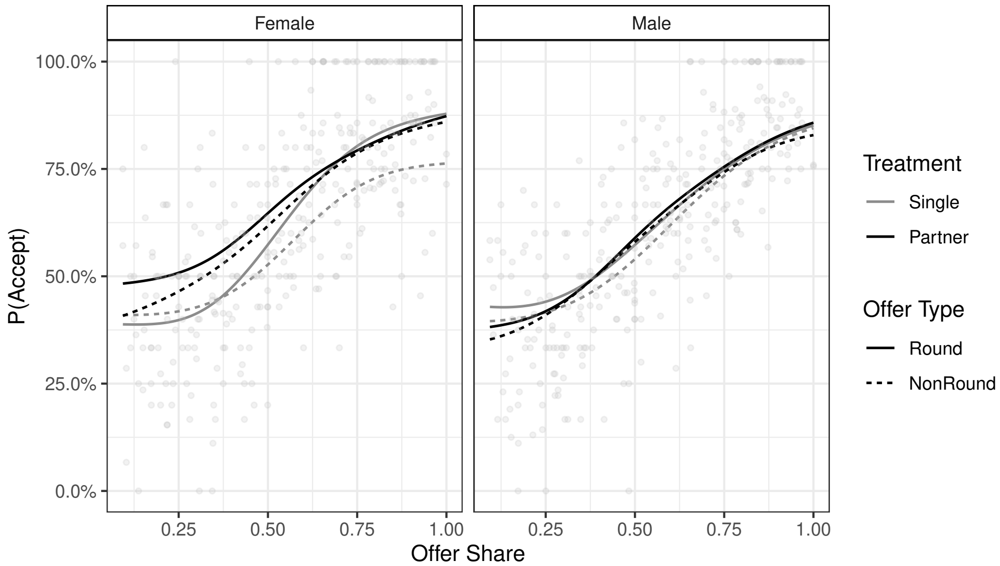


Figure 4.6. Acceptance frequencies for the treatments and offer types by gender.

Note: The curves are based on a Nadaraya–Watson kernel estimator using a normal kernel with a bandwidth of 0.4. The gray dots represent the frequencies to accept. The gray lines represent the Single treatment, and the black lines correspond to the Partner treatment. The offer types are illustrated with solid lines for round offers and dashed ones for non-round offers.

4.3.3 Round-number effects in the female sample

We summarize our findings in Result 1 and Result 2.

Result 1 In Single, we find evidence of *round-number bias* for sufficiently high offers. For lower offers, this bias vanishes.

Result 2 Comparing Single and Partner, we find evidence of round offers serving as *focal points* for sufficiently low offers. For higher offers, the usage as focal point vanishes.

We provide evidence for these results by referring to the four panels in Fig. 4.7. As shown before, the effects differ for high or low offer shares. Hence, we define four equally spaced segments of offer shares (S.1, S.2, S.3, S.4). In Fig. 4.7 these segments are marked by vertical dotted lines (at 0.094, 0.320, 0.547, 0.773, 1). All curves are obtained by a Nadaraya–Watson kernel estimator using a normal kernel with a bandwidth of 0.4. The gray dots represent the frequencies to accept for a given offer share. Note that offer share does not start at zero but 100/1067, as 100 is the smallest possible offer in every period. The gray lines represent the Single treatment, and the black lines correspond to the Partner treatment. The offer types are illustrated with solid lines for round offers and dashed ones for non-round offers. At the bottom of Fig. 4.7, the results of χ^2-tests are reported with *** $p<0.01$, ** $p<0.05$, * $p<0.10$. The test is based on the counts of accepted offers in the respective segments for the indicated category.

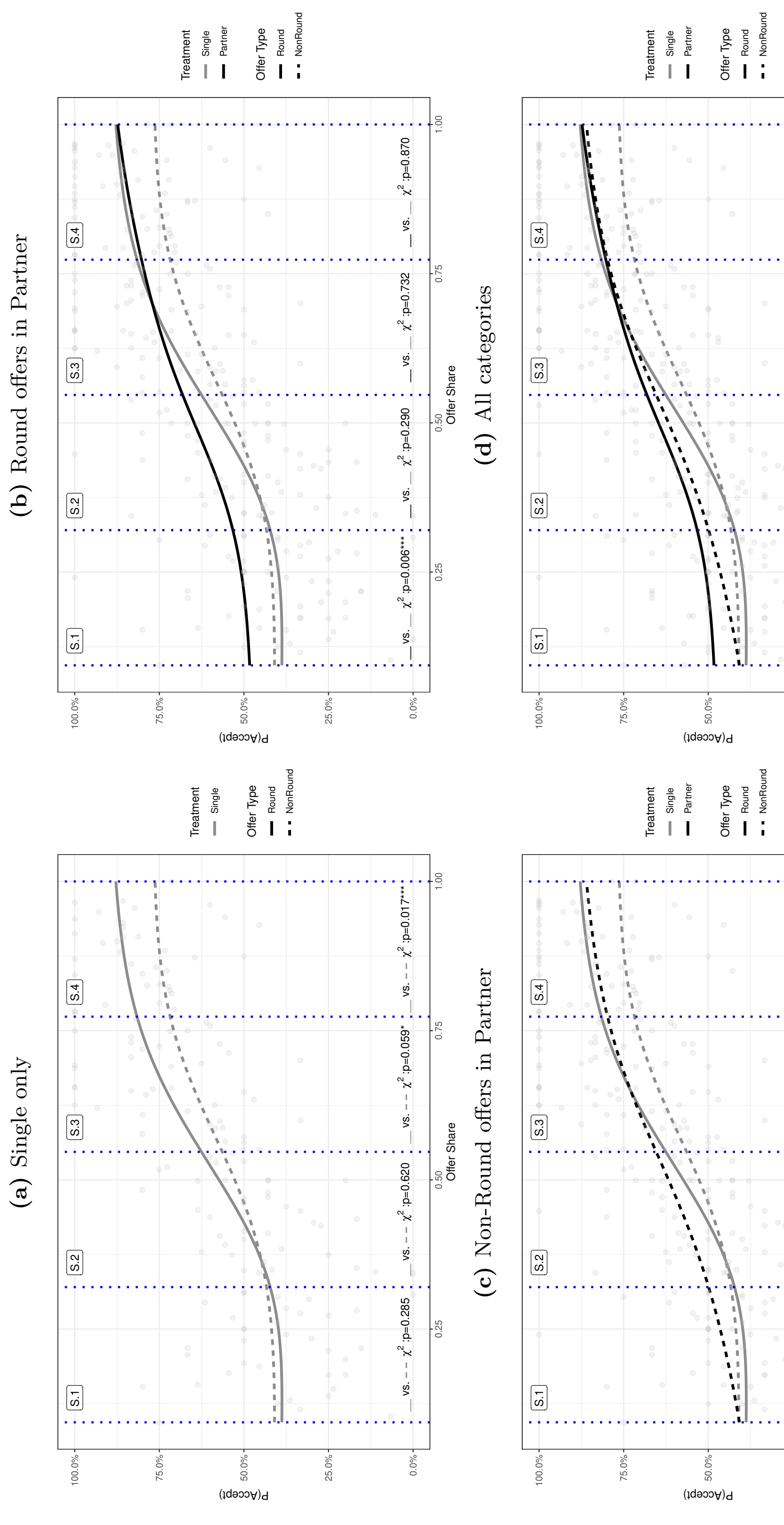


Figure 4.7. Acceptance frequencies for the female sub-sample.

We start by analyzing acceptance frequencies in Single. In this treatment, there is no strategic interaction, and hence decisions reveal pure preferences. The estimated curves can be seen in panel (a) of Fig. 4.7. In S.1, the curve for non-round offers is slightly above the curve for round offers, but the difference is rather small. A χ^2-test shows no significant differences between acceptances at round and non-round numbers in S.1 ($p = 0.285$).[24] In S.2, there seems to be a switching point. Round-number acceptance increases stronger than non-round number acceptance. However, the difference in this segment is also not significant (χ^2, $p = 0.620$). The difference becomes larger in S.3 and S.4. Round-number acceptance is now clearly above non-round number acceptance and significant for these segments (χ^2, S.3: $p = 0.059$, S.4: $p = 0.017$). This is evidence for the presence of round-number effects in Single. As decisions in Single reveal pure preferences, only a *round-number bias* can explain these differences. Thus, we conclude that round-number bias emerges for sufficiently high offer sizes and grows stronger when the offer size increases (**Result 1**). The regression analysis in Appendix B.2.3 confirms this result.

As for the interpretation of these results, we think this hints towards a heuristic decision process where low offers will be declined right away, notwithstanding whether they are round or not. When offer share increases to a certain level, that makes acceptance at least a possibility, this is where the round-number bias comes into play.

In panel (b) of Fig. 4.7, we add the acceptance frequency for round numbers in Partner represented by the black solid line. The acceptance of round numbers in Partner is clearly higher in Single in S.1, and the difference is significant (χ^2, $p = 0.006$). With increasing offer share in the other segments, the difference in the acceptance of round numbers between the treatments vanishes (χ^2, S.2: $p = 0.290$, S.3: $p = 0.732$, S.4: $p = 0.870$). So, when participants have to consider a partner in the Partner treatment, it results in higher acceptance of round numbers for low offer shares.

We now turn our attention to panel (c) of Fig. 4.7. Here, we are able to analyze the treatment effect for non-round numbers. We see that the black dashed line is above the gray dashed line over the entire range of offer shares. This implies that the acceptance frequency for non-round numbers increases when the subjects' payoff also depends on the other player's decision. In S.1, this difference is not significant, with a p-value of $p = 0.995$, as illustrated in the figure. With increasing offer share, the differences become significant (χ^2, S.2: $p = 0.134$, S.3: $p = 0.008$, S.4: $p = 0.009$). Hence, in Partner, acceptance is more likely.

In panel (d) of Fig. 4.7, we add the solid black line indicating the acceptance frequencies in Partner at round numbers. There are two interesting observations concerning this line. First, for high offer shares as in S.3 and S.4, there is no visible difference between round

[24]The χ^2-tests are conducted for each offer share segment. The test hypothesis is at the bottom of each segment. Bar graphs can also represent the frequencies, see Fig. B.1. Since these tests are conducted segment-wise, the test only controls mildly for offer share. The curves serve as a graphical representation of considering the whole range of offer share to compensate for this.

and non-round numbers within Partner. Considering the round-number effects in Single, we would have expected the solid line to be above the dashed line, i.e., again a form of round-number bias as in Result 1. No such bias is present for the Partner treatment in S.3 ($\chi^2, p = 0.354$) and S.4 ($\chi^2, p = 0.557$). Second, for low offer shares, as in S.1, we now find that the black solid line lies above the black dashed line. This difference is significant (χ^2, $p = 0.083$). Thus, there seem to be round-number effects for low offer shares in Partner in segments where we would not expect round-number effects, as none were present in the respective segments in Single. This can evidently not be explained by preferences for round numbers because otherwise, we should have seen round-number effects in Single in S.1. A possible explanation is that round numbers serve a coordinative role that only becomes relevant in Partner (**Result 2**). It might be the case that, once offer share is sufficiently large, this need for coordination becomes less important, as acceptance becomes more likely in general. For increasing offers, we see the pattern previously discussed. In S.3 and especially S.4, there are hardly any differences between the solid black line and black dashed line. This, again, is a striking observation, as it indicates no round-number bias in Partner on segments where we observe a round-number bias in Single. A possible interpretation is that subjects, when confronted with the need for coordination in Partner, shift from heuristic decision-making towards a more thorough computation of expected gains, thereby eliminating the round-number bias. Assuming that subjects operate in such a mode of thorough computation when making decisions in real life, would thus imply that round-number clusters in observational bargaining data are, to a considerable extent, driven by coordination.

To evaluate the robustness of our results, we estimate the linear probability models from Columns (2)-(4) of Table 4.5 separately for each segment. For details, see Appendix B.2.3. The results are qualitatively, quantitatively, and with respect to their statistical significance, in line with our graphical analysis and the non-parametric tests.

4.4 Conclusion

We studied the role of round numbers in bargaining settings. Analyzing observational data, we found that throughout the bargaining process, the share of round-number offers and counteroffers increases. Also, negotiations with a final price that was round were on average shorter, as measured by both, total duration and the length of the offer-counteroffer sequence.

By developing a novel experimental framework, we are able to analyze the differences between round-number effects in individual and cooperative settings. We find robust evidence for the presence of round-number effects in the form of higher acceptance frequencies in our experiment. The channels resulting in these increased acceptance frequencies for round numbers differ between the individual and the cooperative setting, especially for the female sub-sample: Here, in the individual setting, we observe round-number effects only for higher offer shares, while in the coordinative setting, we find round-number effects only for lower offer shares. While the first observation can easily be explained by individual behavioral biases, the latter is apparently the result of coordinative considerations.

Thus, we confirm two possible channels that could induce round-number effects in observational data: (a) individual behavioral biases and (b) round numbers as focal points for coordination in the spirit of Schelling. The observation of round-number effects changing with the offer share points towards a context-dependency of round-number effects. In our experiment, the bias is the main driver for round-number effects when offers are large. For smaller offers, round-number effects are mainly driven by the role of round numbers as focal points.

Our findings conform to a growing body of literature on round-number effects in observational and experimental data. In particular, with a view to the trade-off between saving time and making a better deal, bargaining parties should carefully evaluate the potential impact of the number format and its signaling effect. Hence, using round numbers might be more beneficial in some situations, and in some other scenarios, precise numbers are more useful. Nevertheless, for future experimental research, we advocate taking the number format into consideration when designing studies, as it influences subjects' decision-making.

Chapter 5

On Gender Differences in Competitiveness

Differences in competitiveness have become an essential explanation for labor market outcomes like variations in wages (Card et al., 2016), and different demands in wage negotiations (Leibbrandt and List, 2015). Pinning down the causes and consequences of the willingness to compete is important as it correlates with several relevant choices and characteristics for education and labor market outcomes (Shurchkov and Eckel, 2018). For example, subjects who are more competitive have been found to be more likely to choose competitive educational programs (Almås et al., 2016a; Buser et al., 2014, 2021; Reuben et al., 2017), to have a higher income (Buser et al., 2018a; Kamas and Preston, 2015; Reuben et al., 2015) and to become entrepreneurs (Berge et al., 2015).

During the last decades, an impressive amount of scientific evidence showed that women are generally less competitive than men (Almås et al., 2016b; Balafoutas and Sutter, 2019; Datta Gupta et al., 2013; Niederle, 2017; Niederle and Vesterlund, 2007; Saccardo et al., 2018; Sutter and Glätzle-Rützler, 2015). This *gender gap in competitiveness* (GGC) is robust when using different scientific methods. Studies report that men are more likely to compete when using classical lab (Niederle and Vesterlund, 2007), lab-in-the-field (Gneezy et al., 2009), field (Hogarth et al., 2012), and online experiments (Buser et al., 2021). The findings also replicate when using subjects from different age groups like children (Sutter and Glätzle-Rützler, 2015), students (Niederle and Vesterlund, 2007), and non-students (Andersen et al., 2013).

This chapter serves as a gentle introduction to the literature on gender differences in competitiveness and aims to smooth the transition to Chapter 6. To this end, it presents the preliminary results of an empirical analysis of experiments that have addressed gender differences in competitiveness.

The following consists of a brief introduction of the approach in Section 5.1. Section 5.2 presents the results. Section 5.3 concludes.

5.1 Methodology

The seminal paper Niederle and Vesterlund (2007) is the first to experimentally study the choices of women and men in a competitive setting. Their design (NV design) consists of 3 stages, where participants complete the same real-effort task in each stage under different incentive schemes. Participants first complete the task under piece-rate incentives (Stage 1), i.e., being paid for each correct answer in task, and then under tournament incentives (Stage 2), i.e., only being paid if the participant is the best within a group. In Stage 3, the participants have to choose if their performance in this stage is paid based on piece-rate incentives, like in Stage 1, or according to a tournament outcome, like in Stage 2. Whenever a participant decides on the tournament incentives in Stage 3, s/he is classified as competitive.[25] The participants get to know how many points they collected but receive no feedback on others' performance or how they performed compared to the other participants until the end of the experiment. For a detailed example of the NV design, refer to Chapter 6 and the instructions in Appendix C.11.3.

In the course of the literature review for this chapter, studies were collected that were as similar as possible to the original NV design. The data set builds on the work of Dariel et al. (2017) and Klege et al. (2021).[26]

The data set was collected according to the following criteria. Relevant search portals were used to search for papers citing NV design. Furthermore, various review articles, such as Niederle (2016), were used as a starting point for further searches. The articles already recorded were reviewed in detail. Finally, any cross-reference found during the implementation of the third project, which is presented in Chapter 6, was followed up. Of course, it must be acknowledged at this point that the literature review is not fully comprehensive, and various systematic approaches are still missing.

The data analysis approach is a first attempt at a meta-analysis. For this purpose, the following part is based on excellent introductory works (see Cuijpers, 2016; Higgins et al., 2019; Schwarzer et al., 2015). The data set collects information on the country where the experiment was conducted, on the task that was used, and the type of population (students, adults, children). Additionally, it summarizes the number of male and female participants in the experiment and the number that opted for the tournament incentives. Moreover, it holds details on the group size and whether participants knew the group composition. Besides that, information on the payment and duration are also included.

In research on GGC in the NV design, generally, the main outcome variable is a binary variable with a value of 1 if a participant opted for the tournament incentives and 0 if not. In meta-analyses, it is common practice to use *odds ratios* (OR) or *risk ratios* (RR) to describe the effect size when the outcome is dichotomous. As the odds ratio is less intuitive and often erroneously interpreted, following Higgins et al. (2019)s' suggestion the RR is used.[27] To this end, the *competition ratio* (CR) is introduced. Let the probability to

[25]There is a strand of the literature that uses the change in performance as a measure of competitiveness.

[26]The collected data are online availabe at `Link` or see Online (2022i).

[27]RRs can easily be converted to ORs and vice versa. Let P_c be the probability of the outcome of interest in the control group, then $RR = \frac{OR}{1-P_c+P_c \cdot OR}$ or $OR = \frac{(1-P_c)RR}{1-P_c \cdot RR}$. This nonlinear relationship between OR and RR illustrates that both measures should not be mistaken for each other. The formula above might be useful for the interpretation of logistic regressions (Zhang and Yu, 1998).

compete for women be given by

$$P_i^w = \frac{n_i^{\text{competing women}}}{n_i^{\text{women}}}, \tag{5.1}$$

and for men by

$$P_i^m = \frac{n_i^{\text{competing men}}}{n_i^{\text{men}}}, \tag{5.2}$$

where $n_i^{\text{competing women}}$ collects the women that opted for the tournament incentives in Stage 3, and n_i^{women} is the number of all women in study i. The same numbers are collected for the men. Then, the competition ratio for study i is given by

$$CR_i = \frac{P_i^m}{P_i^w}. \tag{5.3}$$

For non-zero probabilities P_i^w and P_i^m, the ratio CR_i ranges from 0 to infinity. The case $CR_i > 1$ represents that men are more likely to enter competition than women, and the case $CR_i < 1$ represents the counterfactual.

The data of Niederle and Vesterlund (2007) can serve as an example. The sample size is 80. There are $n^{\text{women}} = 40$ women and $n^{\text{men}} = 40$ men. Of all the men, $n^{\text{competing men}} = 29$ opted for the tournament incentives, that is $P^m = 72.5\%$. Of all the women, $n^{\text{competing women}} = 14$ opted for the tournament incentives, that is $P^w = 35.0\%$. Hence, the competition ratio is $CR = 2.07$, and men are more likely to choose the tournament scheme.

Since a considerable between-study heterogeneity is anticipated due to the differences in tasks, countries, and research questions, a random-effects model was used to pool effect sizes (Schwarzer et al., 2015, Ch. 3.4). Moreover, it cannot be assumed that the studies are exact replications of the original design. In this case, it is conventional to use the random-effects model. This model allows for a distribution of true effect sizes and does not assume that there is only one true effect size. The Knapp-Hartung adjustment (Knapp and Hartung, 2003) is used to calculate the confidence interval around the pooled effect for more adequate error rates (IntHout et al., 2014). The between-study heterogeneity, τ^2, was estimated by the Mantel-Haenszel method (Mantel and Haenszel, 1959) without continuity corrections for risk ratios (Greenland and Robins, 1985).

5.2 Results

The collected studies are summarized by a *forest plot.*[28] Fig. 5.1 presents 47 studies and lists them by their subject pool (student, adult, or child), country, and the real-effort task. If one study reports different samples, each sample forms a row in the forest plot. The column *Comp.* displays the number of N who opt for the tournament payment scheme for men and women. The vertical solid black line is a reference line that marks the no-effect ratio of $CR = 1$. The x-axis uses a logarithmic scale. Column 95% CI shows the lower

[28] The list of references can be found in Table 5.2 at the end of this chapter.

and upper limit of the 95% confidence interval. The column *weight* collects the shares used for the pooling in the random-effects model and is also illustrated in the column *Competition ratio* by the gray square surrounding the point estimate — the square scales with the weight size. The short red horizontal lines show the prediction intervals of the subgroups. At the bottom, the diamond-shaped point denotes the average weighted effect size of the subgroup, and its width represents the confidence interval of the pooled effect.

All CRs can be calculated since the column Comp. for men and women is never zero. It already stands out at first glance that most of the studies' CRs are located on the right side of the reference line implicating that men compete more than women in the NV design.

The random-effects model estimates a pooled effect size of $CR = 1.51$ (95%CI: [1.39; 1.63]). The pooled effect is significant ($t = 10.64$, $p < 0.001$) and indicates that men are 1.5 times more likely to enter the tournament payment scheme than women. The moderate to large heterogeneity ($I^2 = 33\%, 46\%, 80\%$) asks for a more detailed look. The between-study heterogeneity variance was estimated at $\hat{\tau}^2 = 0.055$ (95%CI: [0.0298; 0.1007]), with an I^2 value of 70% (95%CI: [60.6%; 76.4%]). The prediction interval for CR ranged from 0.94 to 2.42, indicating that the reversed competition ratio ($CR < 1$) cannot be ruled out for future studies.

Harrer et al. (2019, Ch. 5.4.1) propose a simple approach to detect outliers: if a study's confidence interval does not overlap with the confidence interval of the pooled effect, it is considered an outlier. Following this approach yields a pooled effect estimate of $CR = 1.55$ (95%CI: [1.4597; 1.6514]) and a substantial reduction of the heterogeneity variance, $\hat{\tau}^2$, to 0.014. The I^2 value is more than halved to 30.7% and the prediction interval for CR is tighter ranging from 1.21 to 1.99. This indicates that men can be expected to compete more than women in future studies. In total, nine entries were removed. The results of the random-effects model approach with and without outliers are summarized in Table 5.1. Moreover, it includes the results of two unsupervised machine learning algorithms – the *k-means algorithm* (Hartigan and Wong, 1979) and the *gaussian mixture models* (GMM) (Fraley and Raftery, 2002) – which yield similar results.

5.3 Concluding remarks

This chapter served as a short introduction to the gender differences literature. Furthermore, it presented a newly compiled data set that summarizes various studies on GGC. Based on this data set, a first meta-analysis was performed, and a competition ratio of approximately 1.5 was estimated. Thus, it can be assumed that men tend to be more willing to choose a tournament payment scheme in this setting.

However, recently some evidence has been collected on the lack of a GGC under certain circumstances. For example, for the matriarchy of Masai in Kenya, adult women are reported to be even more competitive than men (Gneezy et al., 2009). Similarly, children living in the Khasi matrilineal society in northeast India are equally competitive (Andersen et al., 2013). Without the need to go afar, it has been shown that the type of school

Study	Country	Task	Men Comp.	Men N	Women Comp.	Women N	CR	95% CI	weight
Subgroup = Student									
Apicella et al. (2017)	USA	Addition	30	52	18	48	1.54	[1.00; 2.37]	1.4%
Balafoutas et al. (2012)	Austria	Addition	38	65	21	67	1.87	[1.24; 2.81]	1.5%
Balafoutas, Sutter (2012)	Austria	Addition	23	36	11	36	2.09	[1.21; 3.63]	1.1%
Berlin, Dargnies (2016)	France	Addition	73	114	41	114	1.78	[1.34; 2.36]	2.0%
Bönte et al. (2017)	Germany	Quiz	40	71	16	115	4.05	[2.46; 6.67]	1.2%
Brandts et al. (2015)	Spain	Addition	33	56	17	56	1.94	[1.23; 3.05]	1.4%
Buser et al. (2017)	USA	Addition	26	50	15	54	1.87	[1.13; 3.10]	1.2%
Buser et al. (2018)	Denmark	Mix	58	137	42	160	1.61	[1.17; 2.23]	1.8%
Cadsby et al. (2013)	Canada	Addition	8	22	2	22	4.00	[0.95; 16.75]	0.3%
Carpenter et al. (2018)	USA	Addition	30	59	16	53	1.68	[1.04; 2.72]	1.3%
Cason et al. (2010)	USA	Addition	25	44	11	25	1.29	[0.77; 2.15]	1.2%
Dargnies (2012)	France	Addition	33	39	19	37	1.65	[1.17; 2.32]	1.7%
Dariel et al. (2017)	UAE	Addition	33	66	44	81	0.92	[0.67; 1.26]	1.8%
Datta Gupta et al. (2013)	France	Mazes	39	70	22	70	1.77	[1.18; 2.66]	1.5%
Halko, Saaksvuori (2017)	Finland	Addition	29	39	22	41	1.39	[0.99; 1.94]	1.7%
Healy, Pate (2011)	USA	Addition	26	32	9	32	2.89	[1.62; 5.15]	1.0%
Kamas, Preston (2012)	USA	Addition	64	156	36	154	1.75	[1.25; 2.47]	1.7%
Klinowski (2019)	Chile	Addition	16	34	8	45	2.65	[1.29; 5.45]	0.8%
Klinowski (2019)	Chile	Verbal	17	34	9	40	2.22	[1.14; 4.33]	0.9%
Mueller, Schwieren (2012)	Germany	Addition	24	57	18	70	1.64	[0.99; 2.70]	1.2%
Niederle et al. (2013)	USA	Addition	31	42	13	42	2.38	[1.47; 3.88]	1.3%
Niederle, Vesterlund (2007)	USA	Addition	29	40	14	40	2.07	[1.30; 3.29]	1.3%
Price (2012)	USA	Addition	106	160	74	150	1.34	[1.10; 1.63]	2.3%
Reuben et al. (2017)	USA	Addition	48	89	45	168	2.01	[1.47; 2.76]	1.8%
Shurchkov (2012)	USA	Verbal	24	54	10	54	2.40	[1.27; 4.52]	0.9%
Shurchkov (2012)	USA	Math puzzle	16	42	13	42	1.23	[0.68; 2.23]	1.0%
Wozniak et al. (2014)	USA	Addition	50	91	27	89	1.81	[1.26; 2.61]	1.6%
Wozniak et al. (2014)	USA	Verbal	43	82	27	83	1.61	[1.11; 2.34]	1.6%
Zhong et al. (2018)	Singapore	Addition	39	98	32	99	1.23	[0.85; 1.79]	1.6%
Random effects model				1931		2087	1.73	[1.55; 1.93]	40.4%
Prediction interval			.	.	.	.		[1.16; 2.57]	--
Heterogeneity: $I^2 = 46\%$, $\tau^2 = 0.0345$, $p < 0.01$									
Subgroup = Adult									
Apicella et al. (2017)	Mturk	Captcha−Style	52	130	32	115	1.44	[1.00; 2.06]	1.7%
Apicella, Dreber (2015)	Tanzania	Skipping rope	46	102	27	89	1.49	[1.02; 2.18]	1.6%
Apicella, Dreber (2015)	Tanzania	Bead collection	26	50	14	38	1.41	[0.86; 2.31]	1.2%
Apicella, Dreber (2015)	Tanzania	Handgrip strength	26	39	9	31	2.30	[1.27; 4.16]	1.0%
Banerjee et al. (2018)	India	Memory task	19	88	13	80	1.33	[0.70; 2.51]	0.9%
Bönte et al. (2018)	Germany	Math	56	101	56	124	1.23	[0.95; 1.59]	2.0%
Booth et al. (2019)	China	Addition	59	159	57	175	1.14	[0.85; 1.53]	1.9%
Booth et al. (2019)	Taiwan	Addition	94	175	89	184	1.11	[0.91; 1.36]	2.3%
Buser et al. (2018)	Denmark	Matrix task	30	80	26	112	1.62	[1.04; 2.51]	1.4%
Cassar et al. (2016)	China	Addition	62	173	48	185	1.38	[1.01; 1.89]	1.8%
Gneezy et al. (2009)	India (matrilineal)	Bucket toss	11	28	28	52	0.73	[0.43; 1.23]	1.2%
Gneezy et al. (2009)	Tanzania (patriarchal)	Bucket toss	20	40	9	34	1.89	[1.00; 3.58]	0.9%
Klege, Visser (2021)	Rwanda	Addition	94	191	79	183	1.14	[0.91; 1.42]	2.2%
Mayr et al. (2012)	USA	Math	157	281	94	262	1.56	[1.28; 1.89]	2.3%
Random effects model				1637		1664	1.32	[1.16; 1.49]	22.4%
Prediction interval			.	.	.	.		[0.98; 1.76]	--
Heterogeneity: $I^2 = 33\%$, $\tau^2 = 0.0146$, $p = 0.11$									
Subgroup = Child									
Almås et al. (2016)	Norway	Addition	129	250	75	233	1.60	[1.28; 2.00]	2.2%
Andersen et al. (2013)	India (matrilineal)	Bucket toss	50	96	37	76	1.07	[0.79; 1.44]	1.9%
Andersen et al. (2013)	India (patriarchal)	Bucket toss	36	71	29	75	1.31	[0.91; 1.89]	1.6%
Buser et al. (2014)	Netherlands	Addition	87	177	43	185	2.11	[1.56; 2.86]	1.9%
Buser et al. (2017)	Switzerland	Addition	61	90	81	159	1.33	[1.08; 1.64]	2.2%
Cardenas et al. (2012)	Colombia	Math search	57	163	49	152	1.08	[0.79; 1.48]	1.8%
Cardenas et al. (2012)	Colombia	Word search	43	164	44	152	0.91	[0.63; 1.30]	1.7%
Cardenas et al. (2012)	Sweden	Math search	70	158	28	146	2.31	[1.59; 3.36]	1.6%
Cardenas et al. (2012)	Sweden	Word search	62	159	39	146	1.46	[1.05; 2.03]	1.8%
Dreber et al. (2014)	Sweden	Addition	38	106	19	109	2.06	[1.27; 3.33]	1.3%
Dreber et al. (2014)	Sweden	Word search	35	105	31	109	1.17	[0.78; 1.75]	1.5%
Khachatryan et al. (2015)	Armenia	Addition	210	388	196	377	1.04	[0.91; 1.19]	2.5%
Khachatryan et al. (2015)	Armenia	Word search	221	388	211	377	1.02	[0.90; 1.15]	2.5%
Lee et al. (2014)	South Korea	Addition	76	254	86	386	1.34	[1.03; 1.75]	2.0%
Samak (2013)	USA	Fishing task	46	60	52	63	0.93	[0.78; 1.11]	2.3%
Sutter et al. (2016)	Austria	Addition	50	123	24	123	2.08	[1.37; 3.16]	1.5%
Sutter, Glätzle−Rützler (2015)	Austria	Addition	120	298	78	419	2.16	[1.70; 2.76]	2.1%
Zhang (2019)	China (Han)	Addition	30	48	23	48	1.30	[0.90; 1.88]	1.6%
Zhang (2019)	China (Yi)	Addition	29	48	18	48	1.61	[1.05; 2.48]	1.4%
Zhang (2019)	China (Mosuo)	Addition	30	40	19	40	1.58	[1.09; 2.29]	1.6%
Random effects model				3186		3423	1.38	[1.20; 1.59]	37.2%
Prediction interval			.	.	.	.		[0.80; 2.40]	--
Heterogeneity: $I^2 = 80\%$, $\tau^2 = 0.0642$, $p < 0.01$									
Random effects model				**6754**		**7174**	**1.51**	**[1.39; 1.63]**	**100.0%**
Prediction interval								**[0.94; 2.42]**	
Heterogeneity: $I^2 = 70\%$, $\tau^2 = 0.0548$, $p < 0.01$									

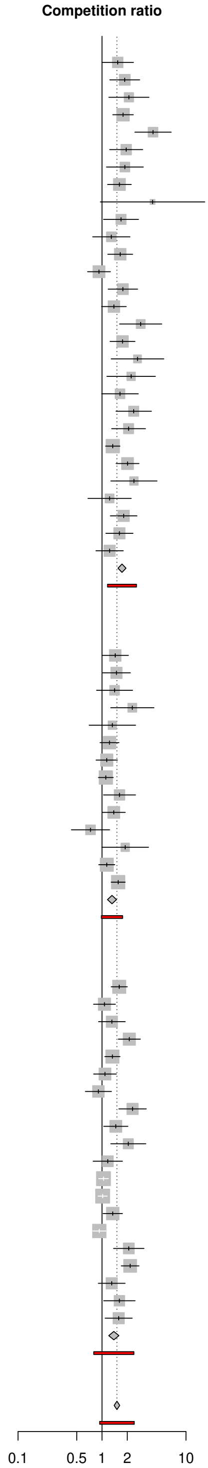


Figure 5.1. Forest plot.

Table 5.1. Heterogeneity analysis of CR.

	Pooled CR	95%CI	p	95%CI	I^2
Main Analysis	1.51	[1.39, 1.63]	<0.001	[0.94, 2.42]	69.5
Outlier removed[a]	1.55	[1.46, 1.65]	<0.001	[1.21, 1.99]	30.7
k-means[b]	1.54	[1.45, 1.64]	<0.001	[1.18, 2.02]	39.3
GMM[c]	1.56	[1.46, 1.66]	<0.001	[1.17, 2.07]	41.4

Note: The table summarizes the pooled CR for three apporaches to reduce heterogeneity. The first row shows the analysis without any study removed.

[a] 9 studies removed: Dariel et al. (2017), Bönte et al. (2017), Booth et al. (2019), Gneezy et al. (2009), Khachatryan et al. (2015): Addition, Word search, Sutter, Glätzle-Rützler (2015), Cardenas et al. (2012), Samak (2013).

[b] 8 studies removed: Dariel et al. (2017), Healy, Pate (2011), Bönte et al. (2017), Gneezy et al. (2009), Khachatryan et al. (2015): Addition, Word search, Cardenas et al. (2012), Samak (2013).

[c] 7 studies removed: Dariel et al. (2017) , Bönte et al. (2017), Gneezy et al. (2009), Khachatryan et al. (2015): Addition, Word search, Cardenas et al. (2012), Samak (2013).

children attend influences competitiveness with female students from girl's schools being as competitive as boys (Booth and Nolen, 2012). Moreover, for children from families with lower socioeconomic backgrounds, no GGC is reported (Almås et al., 2016b). Also, cultural differences play a role in competitiveness, as shown by Cárdenas et al. (2015). They found that children are equally competitive in Columbia, but boys in Sweden are more competitive than girls. These mentioned studies suggest that women's lower willingness to compete is not something that they are born with, but rather a behavioral preference that can be influenced by different factors and can thus be addressed to nurture rather than nature.

Support for this perspective is provided by research showing that the GGC can be closed or reversed when using interventions, which do not influence participants' biological makeup. For example, some studies change the institutional environment to resemble different affirmative action policies and obtain gender balance in competitive environments (Balafoutas and Sutter, 2012; Baldiga and Coffman, 2018; Leibbrandt et al., 2018; Niederle and Vesterlund, 2007). Others use the easy-to-implement intervention of priming (Balafoutas et al., 2018; Cadsby et al., 2013) which encourages women to enter competitions more often. Moreover, giving feedback about relative performance (Wozniak et al., 2016) and the earnings implications related to competition avoidance (Kessel et al., 2021) successfully increases women's entry rates, as well as when more experienced people advise strong-performing women to compete (Brandts et al., 2015). Besides, when the price of the competition benefits not the participants themselves, but their offspring, again no GGC has been observed (Cassar et al., 2016).

However, it is also plausible that biological factors like genes and hormones may lead to different decisions of women and men and are also a primary driver of behavior. Thus, a new and still developing field of research focuses on competitiveness from a more elementary perspective by taking hormones into account. Up to now, there is only one study by Ranehill et al. (2018), which causally analyses the effect of estrogen and progestin (by administrating oral contraceptives) on competitiveness. The authors find no impact of the two hormones on the willingness to compete. All other studies use self-reported hormonal measures by asking female participants about their menstrual cycle day and taking hormonal contraceptives to infer their hormonal level. Using self-reports is noisy (for a detailed discussion why this is the case, see Dreber and Johannesson (2018)) and leads to mixed findings whether hormones play a role for competitiveness or not (Buser et al., 2018a; Wozniak et al., 2014).

The existing evidence already provides results on what factors correlate with competitive behavior and how differences in competitiveness between men and women can be closed. However, the following chapter will be the first to test the robustness of the GGC when priming subjects with a specific gender identity.

Table 5.2. References for the forest plot.

Population	Reference	Country	Task	Sample size
Student	Apicella et al. (2017)	USA	Addition	100
	Balafoutas et al. (2012)	Austria	Addition	132
	Balafoutas et al. (2012)	Austria	Addition	72
	Berlin and Dargnies (2016)	France	Addition	228
	Bönte et al. (2017)	Germany	Quiz	186
	Brandts et al. (2015)	Spain	Addition	112
	Buser et al. (2017a)	USA	Addition	104
	Buser et al. (2018b)	Denmark	Mix	297
	Cadsby et al. (2013)	Canada	Addition	44
	Carpenter et al. (2018)	USA	Addition	112
	Cason et al. (2010)	USA	Addition	69
	Dargnies (2012)	France	Addition	76
	Dariel et al. (2017)	UAE	Addition	147
	Datta Gupta et al. (2013)	France	Mazes	140
	Halko and Sääksvuori (2017)	Finland	Addition	80
	Healy and Pate (2011)	USA	Addition	64
	Kamas and Preston (2012)	USA	Addition	310
	Klinowski (2019)	Chile	Addition	79
	Klinowski (2019)	Chile	Verbal	74
	Müller and Schwieren (2012)	Germany	Addition	127
	Niederle and Vesterlund (2007)	USA	Addition	80
	Niederle et al. (2013)	USA	Addition	84
	Price (2012)	USA	Addition	310
	Reuben et al. (2017)	USA	Addition	257
	Shurchkov (2012)	USA	Verbal	108
	Shurchkov (2012)	USA	Math puzzle	84
	Wozniak et al. (2014)	USA	Addition	180
	Wozniak et al. (2014)	USA	Verbal	165
	Zhong et al. (2018)	Singapore	Addition	197
Adult	Apicella et al. (2015)	Tanzania	Skipping rope	191
	Apicella et al. (2015)	Tanzania	Bead collection	88
	Apicella et al. (2015)	Tanzania	Handgrip strength	70
	Apicella et al. (2017)	Mturk	Captcha-Style	245
	Banerjee et al. (2018)	India	Memory task	168
	Bönte et al. (2018)	Germany	Math	225
	Booth et al. (2019)	China	Addition	334
	Booth et al. (2019)	Taiwan	Addition	359
	Buser et al. (2018a)	Denmark	Matrix task	192
	Cassar et al. (2016)	China	Addition	358
	Gneezy et al. (2009)	India (matrilineal)	Bucket toss	80
	Gneezy et al. (2009)	Tanzania (patriarchal)	Bucket toss	74
	Klege et al. (2021)	Rwanda	Addition	374
	Mayr et al. (2012)	USA	Math	543
Child	Almås et al. (2016b)	Norway	Addition	483
	Andersen et al. (2013)	India (matrilineal)	Bucket toss	172
	Andersen et al. (2013)	India (patriarchal)	Bucket toss	146
	Buser et al. (2014)	Netherlands	Addition	362
	Buser et al. (2017b)	Switzerland	Addition	249
	Cárdenas et al. (2012)	Colombia	Math search	315
	Cárdenas et al. (2012)	Colombia	Word search	316
	Cárdenas et al. (2012)	Sweden	Math search	304
	Cárdenas et al. (2012)	Sweden	Word search	305
	Dreber et al. (2014)	Sweden	Addition	215
	Dreber et al. (2014)	Sweden	Word search	214
	Khachatryan et al. (2015)	Armenia	Addition	765
	Khachatryan et al. (2015)	Armenia	Word search	765
	Lee et al. (2014)	South Korea	Addition	640
	Samak (2013)	USA	Fishing task	123
	Sutter and Glätzle-Rützler (2015)	Austria	Addition	717
	Sutter et al. (2016)	Austria	Addition	246
	Zhang (2019)	China (Han)	Addition	96
	Zhang (2019)	China (Yi)	Addition	96
	Zhang (2019)	China (Mosuo)	Addition	80

Chapter 6

Decisions have no Gender. Gender and Economic Decision-making revisited[29]

Worldwide, humans make economic decisions every day: Should I apply for a new job opportunity in a highly competitive environment? Should I invest in a risky asset or not? How much money should I donate to charities? A vast literature tries to determine the factors that affect decisions in domains such as competitiveness (Villeval, 2012), risk–taking (Thöni and Volk, 2021), and altruism (Bilén et al., 2021). Researchers have looked, amongst others, into the role of institutional or marketrelated features (Balafoutas et al., 2018; Balafoutas and Sutter, 2012; Cassar and Rigdon, 2021; Cassar et al., 2016; Fornwagner et al., 2022b; He et al., 2021; Niederle and Vesterlund, 2007), cultural background (Cárdenas et al., 2015; Croson and Gneezy, 2009; Gneezy et al., 2009; Gong and Yang, 2012; Liu and Zuo, 2019), individual characteristics (Almås et al., 2016b; Buser et al., 2018a; Guiso and Paiella, 2008; Sutter and Glätzle-Rützler, 2015; Von Gaudecker et al., 2011), hormonal (Boksem et al., 2013; Ranehill et al., 2018; Sapienza et al., 2009; Van Anders et al., 2015; Zak et al., 2009; Zethraeus et al., 2009), or other biological factors, such as genetics, and neurological factors (Anderson et al., 2015; Cesarini et al., 2012; Moll et al., 2006; Reuter et al., 2011). Among those factors, gender has received a lot of attention. Over the last few decades, the flourishing research in economics has come to the conclusion that gender is a significant driver of how women and men behave: gender differences in behavior are a common finding for competitiveness (Beblo and Markowsky, 2022), risk–taking (Thöni and Volk, 2021), and altruism (Bilén et al., 2021). We refer to Appendix C.10 for a more detailed literature review on risk and altruism.

But do observed data really show gender differences? Is it instead sex differences that influence behavior, or is it a mix of gender and sex? Importantly, sex and gender are two distinct concepts. Whereas sex is defined as "either of the two main categories (male and female) into which humans" are categorized based on their reproductive functions[30], gender

[29] This chapter is based on the joint work Fornwagner et al. (2022a).

[30] See `Link` or Online (2022c).

usually refers to the psychological, behavioral, social, and cultural aspects of being male or female (i.e., masculinity or femininity) (VandenBos, 2007). For cisgender individuals, their internal gender identity matches and presents itself by the externally determined cultural expectations of the behavior and roles considered appropriate for one's sex (VandenBos, 2007). However, the gender identity of transmen and transwomen and their gender roles are not the same as what is typically associated with their sex assigned at birth (American Psychological Association, 2015). So the question arises: how much of the differences of men and women often found in the economic literature can really be associated with gender as opposed to an individual's sex?

We investigate this question by using well–known behavioral economic experiments in the domain of *competitiveness*, *risky choices*, and *altruism*. As stated, for these three behavioral traits, gender differences are a common finding. However, these differences have usually been observed using cismen and ciswomen as subjects, which differ in their gender *and* sex. Distinguishing gender from sex effects is practically impossible when only investigating cisgender participants. As a novel approach, we run our experimental study with transmen and transwomen in addition to cismen and ciswomen. The advantage is that cisgender and transgender people differ in either their sex or their gender. To illustrate this consider an example: a ciswoman has female sex *and* feminine gender. A transman has female sex *but* masculine gender. So differences in the behavior of those two subject groups might be associated with gender instead of sex. The experimental method is excellent for studying the economic choices we are interested in because of its standardized and validated measures. We have information on the participants' gender and sex from self–reported categories and established scaling methods from psychological and medical science. Moreover, instead of just analyzing gender and sex effects correlationally, we elicit the causal impact of gender by exogenously varying gender identities with a priming method.

First, we test how gender correlates with the mentioned choices. By contrasting the behavior of the four different subject groups of cismen, ciswomen, transmen, and transwomen, we obtain insights into how far biology (sex) or the cultural and sociological construct of gender explains differences in economic behavior. Our study is the first investigating competitiveness, risk–taking, and altruism of transmen and transwomen. We hypothesize that if gender is the driving factor, individuals of the same gender (and different sex) make similar decisions, and decisions significantly differ when gender differs (and sex is the same). Second, we concentrate on the causal effect of gender on behavior – an analysis that is rarely done in the literature. The traditional experimental method of randomizing over the variable of interest is not possible with gender. Hence, we need a different approach to elicit causal effects. As our method to test a directional impact of gender, we employ a gender prime: either a masculine or feminine gender identity is subconsciously activated. Priming is a very powerful, easy–to–implement intervention to activate gender identities (Rudman and Phelan, 2010; Steele and Ambady, 2006). If cisgender and transgender individuals change their behavior when being primed, it indicates a causal effect of gender on individual economic decisions.

Based on 780 observations from experiments conducted online, our results generally show no correlational or causal effect of gender or sex for competitiveness, risk-taking, and altruism. The only exceptions are that cismen have a higher rate of entering the competition than all other subject groups when primed masculine. Besides, we find that subjects of male sex (i.e., cismen and transwomen) risk more than their female counterparts (ciswomen and transmen). Moreover, cismen risk more when primed with a masculine identity compared to the neutral priming condition. Thus, in general, we conclude that gender is not a consistent main factor influencing the economic decisions measured in this article.

The remainder is structured as follows. Section 6.1 presents the methodological framework. In Section 6.2, we present our results. Section 6.3 concludes.

6.1 Methods

To test our research questions, we set up an online economic experiment. We conduct our study (tasks and questionnaires) with oTree (Chen et al., 2016) on Prolific (see Online, 2022j). Each participant completes six parts and several questionnaires. One part is randomly selected for payment at the end of the experiment. In Part 1, a participant is randomly assigned to either the baseline treatment (NEUTRAL) or a treatment condition that refers to one of the gender priming interventions: FEMININE (primes a feminine gender identity) or MASCULINE (primes a masculine gender identity). Participants are primed by a word search task where different words are used depending on the underlying treatment (Bargh et al., 2001). The words in FEMININE are: female, woman, she, women, her, girl, hers, lady; in MASCULINE, they are: male, man, he, men, him, boy, his, gentleman. In the baseline condition NEUTRAL, participants also solve the word search task, with the following (neutral) words: person, it, people, its, child, theirs, individual, neuter. Participants are shown the words and have two minutes to mark these words in a 10×10 grid. In case they find all words, they receive £5.

After the word search task, each participant enters the next parts of the experiments, which are the respective economic decision–making parts. As our first decision dimension, we employ monetary incentives to measure competitiveness (Buser et al., 2021). We measure performance in a real effort math task, where the participants are instructed to solve puzzles by finding two two–digit numbers that add up to 100 in 3×3 matrices for two minutes. In Part 2, they complete the math task under piece–rate incentives, which means they receive £0.50 for every solved puzzle. In Part 3, the same math task is performed under tournament incentives. The participants are divided into groups of four and receive £2 for every solved puzzle, but only if they solve more puzzles than every other group member. In Part 4, the participants have to choose, before performing, if their performance in this part will be paid based on the piece–rate incentives (like Part 2) or according to the tournament rules (like Part 3). Whenever a participant decides on the tournament incentives in Part 4, s/he is classified as competitive and competes against the group member's performance in the previous Part 3. In all parts, the participants do not receive feedback on how well they perform compared to the other group members until the end of the experiment and have no information on the other group members' identity or characteristics. Additionally, we measure the participants' confidence in Part 2 (how well they think they performed compared to the other participants in the session) and Part 3 (how well they think they performed compared to the other group members) with incentivized questions.

Our second decision dimension is the willingness to take risks in Part 5. It is measured using a simple lottery task (Gneezy and Potters, 1997). Participants receive £4 and can invest into a lottery with a 50% chance of success. The invested amount is multiplied by 2.5 in case of success. In case of no success, the invested amount is lost. The participants keep the amount not invested. Risk preferences are measured as the amount a participant invests, where higher investments indicate a higher willingness to take risks. The third decision dimension is altruism in Part 6. We investigate participants' altruistic preferences with a dictator game (Kahneman et al., 1986). Participants receive £5 and split up this amount between themselves and up to five different charities. Altruism is quantified as the sum donated by a participant.

The post–experimental questionnaire contains (1) a 30–items version of the Bem Sex Role Inventory (BEM) that explores a person's masculine and feminine self–identification on a continuous scale (Geldenhuys and Bosch, 2020); (2) the Transgender Congruence Scale (TCS) (Kozee et al., 2012) which evaluates if someone identifies as transgender; (3) demographic questions, as well as questions on the biological sex, gender, sexual orientation, and whether one self–identifies as transgender; and (4) the Steps to Transition (STT) questionnaire that describe typical steps transgender people undertake in their transition (Kozee et al., 2012). In addition, we include debriefing questions to check if the participants are aware of the study topic and the priming intervention (Chartrand and Bargh, 1996).

Appendix C.11.3 provides a detailed description of all instructions and questionnaires of our experiment.

6.2 Results

In order to summarize the extensive analysis, we use the following abbreviations for our results: Chi–squared test (χ^2), Kruskal–Wallis test (KW), Kendall's rank correlation coefficients test (KTAU), two–sided Mann–Whitney U test (MWU), Robust Wald test (W), and standard deviation (SD). The significance levels are defined as follows: $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***), where a significant result must have at least $p < 0.05$. We summarize multiple p–values by $p's$.

6.2.1 Descriptives

We collected a total of $n = 780$ observations, out of which 425 are cisgender (214 cismen and 211 ciswomen) and 355 transgender (215 transmen and 140 transwomen; see Appendix C.11 and Appendix C.11.1 for more details). As summarized in Table C.1, the participants are on average 24.4 years old (SD = 6.60), have an average height in centimeters of 170 (SD = 10.8), and approximately half of them are students (47.2%). Around one–third holds a university degree, 69.4% have an income lower than £20,000, and 25.8% report being religious. Our sample consists mostly of participants from the United States, followed by Continental Europe and the United Kingdom. Less than 10% are not residents of the three mentioned regions.

Responses to the BEM classify 28.5% as feminine, 19.4% as masculine, 24.1% as androgynous, and 28.1% as undifferentiated. On the TCS scale ranging from 1 to 5, participants show an average score of 3.67 (SD = 1.1). The average score on the STT, which ranges from 0 to 16, is 4.35 (SD = 4.6). The various subject groups are comparable in several characteristics as indicated by the statistical tests added in Table C.1. Descriptive statistics broken down by subject groups are presented in Tables C.2 and C.3 (cisgender) as well as Tables C.4 and C.5 (transgender).

For the outcomes of Part 1, the detailed Appendix C.2 summarizes descriptives on the participants' priming. On average, the participants marked 7.45 out of 8 words (S.D. = 1.53), and 83.97% (i.e., $n = 655$) marked all words from the list within the given time of two minutes.

6.2.2 Competitiveness

Fig. 6.1 and Table C.14 summarize the tournament entry rates in Part 4. In order to investigate whether gender and competitiveness are correlated, we focus on the baseline treatment NEUTRAL. No significant variation is reported across the four subject groups (χ^2, $p = 0.939$). Similar, when pooling the results by gender (Fig. C.2; cismen + transmen vs. ciswomen + transwomen), tournament entry rates do not differ for feminine and masculine subjects (χ^2, $p = 0.601$) and also no difference is found for male and female subjects when pooling the data by sex (Fig. C.3; cismen + transwomen vs. ciswomen + transmen; χ^2, $p = 0.867$). We compare the differences between the priming conditions (FEMININE and MASCULINE) and the baseline treatment (NEUTRAL) for the causal analysis. Priming does not influence competition entry rates for any subject group (χ^2, $p's > 0.073$). The result is marginally significant only for cismen when comparing the MASCULINE treatment to the NEUTRAL treatment (χ^2, $p = 0.073$). We shall see in the regression analysis that when adding further controls, the impact of MASCULINE priming on cismen becomes significant. Looking at the MASCULINE priming condition only, where the entry rates look very similar for all subject groups except for cismen, the competition entry rate is around 20 percentage points higher for cismen than for all other subject groups (χ^2, $p = 0.046$).

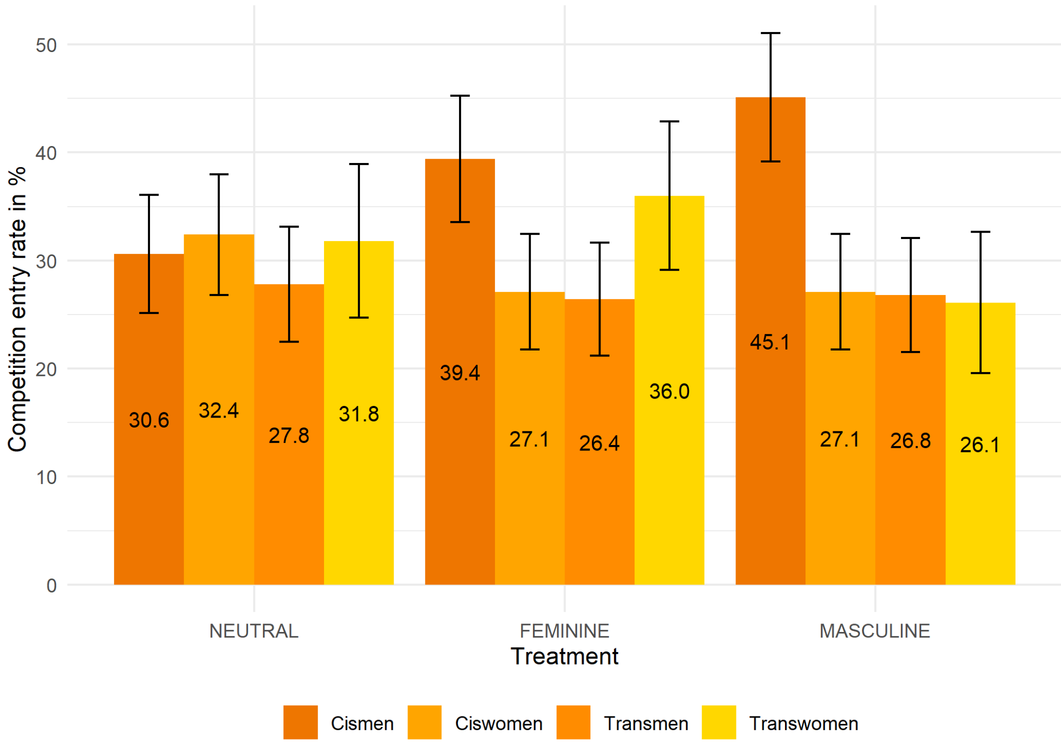


Figure 6.1. Tournament entry rates in Part 4 by treatment and subject group ($n = 780$).

Note: The bars show the percentage of participants (between 0 and 100) who chose to compete rather than to perform under piece–rate incentives. The error bars represent the standard errors of the means.

In Table C.15, we run Probit regressions for the baseline treatment (NEUTRAL) to disentangle the effects of gender and sex. As our basic regression framework, we have in column (1) just the subject groups and in (2) additionally control for the performance measures in the real effort task. In column (3), we further control for confidence and the willingness to take risks. In column (4), we add the variables age, height, student status, income, religion, and residence, whereas in (5), we control for the outcomes in the TCS and STT. Using joint coefficient tests (see Table C.15), we find neither gender (W, $p's > 0.437$) nor sex (W, $p's > 0.214$) to have a significant effect on competitiveness. We thus conclude that there is no correlation between either gender or sex and competitiveness in our study.

To analyze a potential causal effect of gender, we run Probit regressions in Table C.16. The non–parametrized analyses are confirmed for ciswomen, transmen, and transwomen. For cismen we find that the gender prime with MASCULINE has a significant impact increasing the competition entry rates in specification (2) ($p = 0.034$; controlling for performance) and (4) ($p = 0.021$; controlling for beliefs, risk attitude, and other person–specific covariates). Summing up, only cismen's competition entry rates seem to be influenced (positively) when priming them with their own gender identity. We do not find a significant impact of gender priming for all other subject groups and priming combinations. We will interpret those results in the Discussion.

Our experimental design does not only allow us to look into the choice to enter a tournament but also into participants' confidence (i.e., how well they believe they performed in the real effort task when competing, see Table C.11). In NEUTRAL, there is no evidence that subjects of masculine gender have higher performance beliefs than subjects of feminine gender (MWU, $p = 0.362$). However, we do find differences between subjects of female and male sex (MWU, $p = 0.001$). For priming, no subject group increases or decreases their beliefs when being primed (MWU, $p's > 0.177$). Regressions in Table C.12 confirm that beliefs depend on the participants' sex: male subjects generally have higher confidence in their performance than female subjects (W, $p's < 0.001$). And again, confidence does not differ across gender (W, $p's > 0.259$). That gender does not play a role in this setting is further confirmed when looking at the causal impact of gender priming on the participants' confidence. For none of the subject groups, we do find any effect of gender priming on the beliefs when using regression analyses (see Table C.13, W, $p's > 0.178$).[31]

6.2.3 Risk

Investment rates in the lottery are depicted in Fig. 6.2 and stated in Table C.19. When applying non–parametric tests, we do not find any differences between the various subject groups within the baseline treatment NEUTRAL (KW, $p = 0.194$). If anything, transwomen seem to be more risk–prone than transmen in a pairwise comparison (MWU, $p = 0.048$). This, however, does not point towards a systematic impact of gender and/or sex when pooling data (Fig. C.4 and Fig. C.5; gender: cismen + transmen vs. ciswomen + transwomen, sex: cismen + transwomen vs. ciswomen + transmen; MWU, $p's > 0.130$). Turning to the causal impact of priming, again, we see MASCULINE priming increases the risk attitude for cismen only (MWU, $p = 0.038$) bringing the level of cismen to the one of transwomen in the MASCULINE priming (MWU, $p = 0.876$). For every other subject group, we do not find any significant impact of gender priming (MWU, $p > 0.206$).

Joint coefficient tests for the regressions (with and without control variables) in Table C.20 show the correlational results for our baseline condition. We find no differences in risk–taking of subjects of feminine and masculine gender (W, $p's > 0.132$). However, we find a sex effect: male subjects risk more than female subjects (W, $p's < 0.042$).

Turning to priming, we have significant differences in risk–taking of cismen when being primed MASCULINE (W, $p's < 0.046$; see Table C.21). We find no difference in risk–taking for all other subject groups when primed with a gender (W, $p's > 0.092$). The findings are independent of what other control variables are taken into account. The regression analysis for risk attitudes is thus similar to what we found for competition entry rates. There is no systematic influence of a gender prime on the participants. However, when being primed with their own gender, cismen significantly increase their risk taking behavior.

[31] It may be interesting in what payoffs behavior in the competitiveness task results. We provide details and different analyses on the performances in the real effort task of Part 2 to 4 in Appendix C.3.

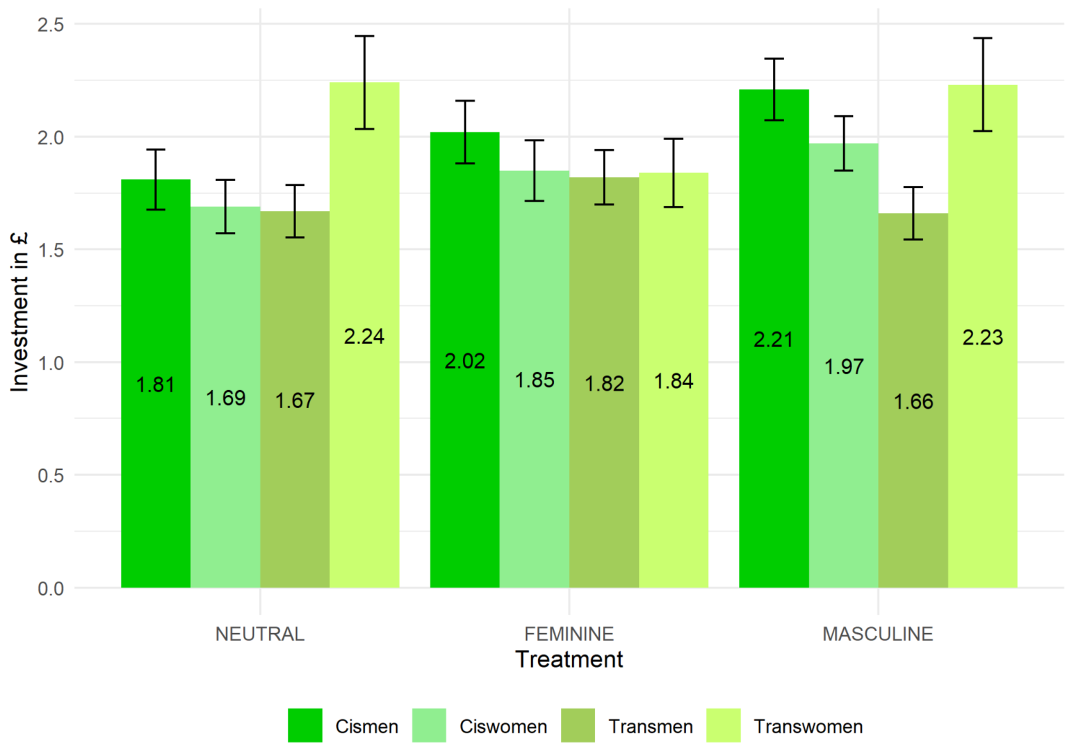


Figure 6.2. Investments into the risky lottery in Part 5 by treatment and subject group ($n = 780$).

Note: The bars show the average investment rate, and the error bars represent the standard errors of the means.

6.2.4 Altruism

Last, we test for differences in the donation task (see Fig. 6.3 and Table C.22). Donations in NEUTRAL are not distinguishable across subject groups (KW, $p = 0.933$). Neither pooled results for gender nor for sex yield a difference in donation rates (Fig. C.6 and Fig. C.7; MWU, $p's > 0.564$). Concerning the causal impact of gender priming, we do not find significant effects for any subject group and any priming condition (MWU, $p's > 0.260$).

The regression analysis in Tables C.23 and C.24 confirms these findings. Joint coefficient tests for gender or sex do not show significant correlations in the baseline condition (W, $p's > 0.580$). Moreover, the impact of all priming conditions on all subject groups remains insignificant, even after controlling for different sets of additional personal covariates (W, $p's > 0.214$).

To summarize, we find no correlation between gender or sex on altruism and do not detect any causal impact of gender priming on altruistic behavior in our setup.

6.2.5 Gender and sex differences within priming conditions

As we have shown so far, there is no systematic correlation between gender and behavior in the NEUTRAL treatment. Here we briefly test for gender and sex differences in behavior within the two priming treatments. Looking at Fig. C.2, Fig. C.3, Fig. C.4, Fig. C.5, Fig. C.6, and Fig. C.7 and analyzing the gender differences with non–parametric tests, we

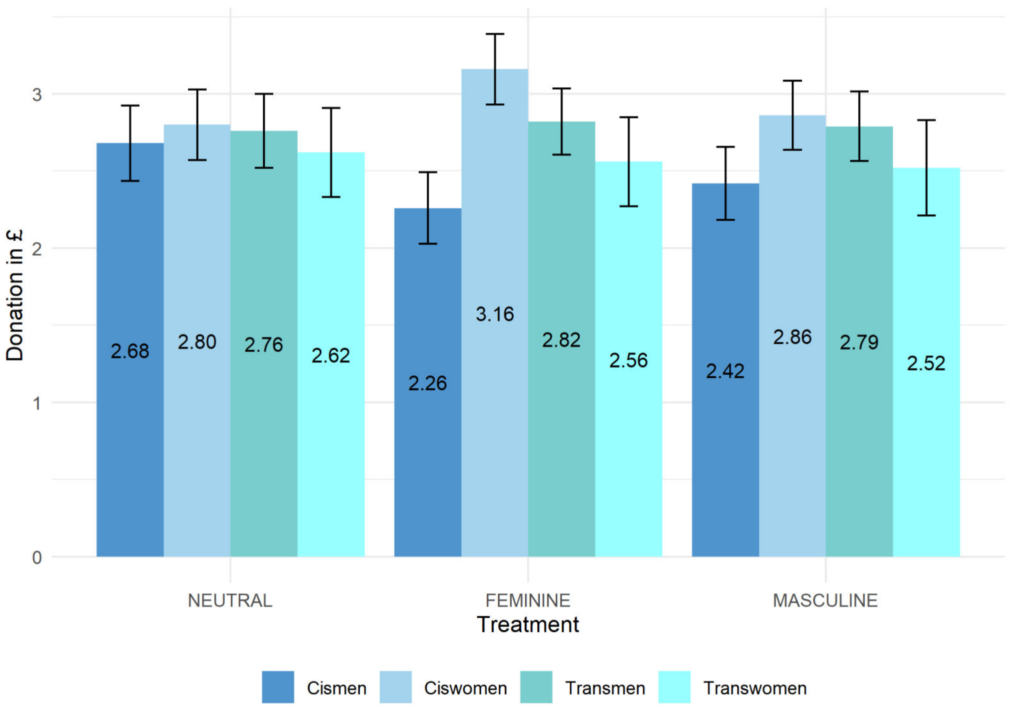


Figure 6.3. Donation in Part 6 by treatment and subject group ($n = 780$).

Note: The average donations are indicated by the bars, and the error bars represent the standard errors of the means.

see no difference in competition entry rates (FEMININE: χ^2, $p = 0.725$, MASCULINE: χ^2, $p = 0.115$), risk–taking (FEMININE: MWU, $p = 0.560$, MASCULINE: MWU, $p = 0.507$), and altruism (FEMININE: MWU, $p = 0.132$, MASCULINE: MWU, $p = 0.532$). Turning to sex differences, the picture slightly changes. First, we see differences between subjects of male and female sex in both priming conditions (FEMININE and MASCULINE) for competitiveness. The differences are close to conventional levels of significance (FEMININE: χ^2, $p = 0.051$, MASCULINE: χ^2, $p = 0.067$). Second, for risk–taking, we find a significant difference in the MASCULINE treatment only, with subjects of male sex taking more risk than subjects of female sex (MWU, $p = 0.011$). Third, for altruism, we find subjects of female sex having significantly higher scores than those of male sex in the FEMININE treatment (MWU, $p = 0.023$). Hence, for risk and altruism we find that only those sexes show higher scores who are primed with the gender identity that they would cisgender–stereotypically be associated with.

6.2.6 Replication of the correlational analysis with a continuous gender measure

With just a handful of exceptions (Kastlunger et al., 2010; Lemaster and Strough, 2014; Meier-Pesti and Penz, 2008), researchers in economics always used a categorical way to measure gender. However, it is more and more discussed that gender might be a continuous characteristic rather than a binary (or categorical) one (Hyde et al., 2019). Gender can be measured on a continuous scale by using the BEM sex role inventory (Geldenhuys and Bosch, 2020), which is part of our post–experimental questionnaire. Thus we rerun all regression analyses and include, instead of the subject groups, the variables $BEMscore : Feminine$ (defined as the score participants reached on the BEM questions measuring femininity) and $BEMscore : Masculine$ (score on masculine questions in the BEM).

Results in Table C.25 to Table C.27 show throughout that neither the feminine nor the masculine score significantly influence how the participants decide ($p's > 0.057$). This is not surprising since the BEM scores and the gender categories are highly correlated (feminine: KTAU, $p = 0.001$, masculine: KTAU, $p = 0.003$), and we did not find correlational gender differences in the baseline condition for neither of the economic decisions we investigate.

6.3 Discussion

This paper applies well–known and extensively used experimental techniques to identify the influence of gender and sex on economic decision–making. First, we separate the impact of gender and sex on economic decisions by collecting data from participants whose gender and sex differ, which is new to the literature. We compare the competitive, risk, and altruistic behavior of four different subject groups – cismen, ciswomen, transmen, and transwomen. Second, we induce either a neutral, feminine, or masculine gender identity by having different priming conditions. Thus, with our experimental setup, we go beyond correlating gender and sex with decisions and try to evoke gender identities through a priming manipulation causally.

Even if this study was pre–registered and carefully designed following existing literature and the state of the art standards in experimental economics, the findings largely diverge from previous work. Our results do not show conclusive correlational or causal evidence for gender or sex as determinants of economic decision–making. Apart from some differences described in the previous sections, the pattern is essentially consistent: gender and sex differences in behavior remain mostly statistically indistinguishable. Moreover, as a side result, we see that cis- and transgender participants do not systematically differ from each other in their behavior. Additionally, the main correlational findings replicate when applying a continuous instead of a categorical gender measure. Our overall interpretation of the data is that gender and sex might not matter as much as we initially thought. But what can explain these findings?

First, one explanation could be that gender effects might depend on the underlying subject pool. The existing literature has treated gender differences in behavior as a pretty well–established and robust finding. However, the vast majority of these papers use standard student subjects (Marianne, 2011). Studies that use other samples (Charness and Villeval, 2009) or online samples are generally less likely to report gender differences, especially when controlling for a set of participants' characteristics (Almås et al., 2016b; Flory et al., 2018)

Second, almost two decades have passed since the first studies that looked into competitiveness, risk, and altruism were published and found gender differences in behavior. One can thus speculate that female empowerment, educational initiatives, and the broader awareness of gender and sex equality in private and professional settings have led to a narrowing of potential behavioral differences in the meantime.

Third, the absence of an effect of gender priming on the behavior of transgender subjects may be rooted in the connotation those subject groups have with gender. For transgender individuals, the concept of gender might be a relatively continuous spectrum, whereas for cis–individuals it might be seen as a binary dimension. As such, gender might not be as decisive for transgender as for cisgender individuals. The fact that gender priming seems to work only for cismen but not for ciswomen might hinge on the role gender usually has played for those two subject groups. Whereas for cismen their gender usually comes with advantages and, as such, has a positive connotation, ciswomen might have negative experiences concerning the way society treats them based on their gender.

Despite the partly unexpected findings, we belief that there are several key "takeaways" from this study. For the first time, we present evidence from a sample of cis- and transgender participants in one framework, which allows for both a correlational and a causal approach, and look at how they decide in a competitive context and when making risky or altruistic decisions. Transgender individuals have become a more and more visible part of society. Thus, we think it is crucial to understand their economic preferences. Besides, having transgender participants in our sample makes it possible to look deeper into the part that an individual's gender - as opposed to sex - plays in economic decision–making. In our setting, we shed light on the part of gender effects that can be attributed to biological factors (which refer to a participant's sex) and other aspects of one's gender identity. Additionally, we do not measure gender only on a categorical scale; instead, we also apply a continuous gender scale. Our results are qualitatively the same, independent of what gender scale is used. Based on our findings, we conclude that the role of gender and sex is not as decisive for economic behavior as previously assumed.

Chapter 7

Conclusion

This dissertation is a collection of three projects that contributed to different research fields. All projects have in common that the approach has been designed from the beginning to be data-driven. The first one addresses curve data that exhibit a complex and challenging structure. The problem of irregularly spaced observations and truncation is solved by a cutting-edge neural network – the RNAA. The second and the third project are based on data collected in online experiments using oTree. They served to analyze the role of round numbers in bargaining and the influence of gender and sex on economic decision–making.

In Chapter 3, the RNAA was introduced. This is a recurrent neural network with encoder-decoder architecture, an attention mechanism, and an imposed autoregressive input data structure. The RNAA learns to match an input sequence to a target sequence and their within-sequential dependencies while autonomously deciding which part of the encoded input sequence should be paid attention to. In an application to data of the German balancing market, the RNAA showed convincing performance in three different evaluation cases. The project contributes a new forecasting method that can operate on curve data that usually need processing for other approaches. The method is accompanied by a toolbox that allows utilization in other applications. Furthermore, the project provides a clean data set for further research in the field of functional time series analysis. Finally, a benchmark study comparing univariate methods, neural networks, and methods from functional time series analysis was also conducted in the context of this project.

The study of the role of round numbers in bargaining settings was presented in Chapter 4. During the analysis of millions of observations from the trading platform eBay, it was found that the share of round-number offers and counteroffers increases throughout the bargaining process. In addition, it was shown that negotiations that ended with a round price were shorter on average. This was confirmed for the duration of the negotiation as well as for the number of offers and counteroffers. In order to study the two possible channels - round-number bias and focal points - an experiment was developed. The experiment was conducted on Amazon MTurk, and it provided robust evidence for a round-number effect that differs between the individual and cooperative setting. For the female sub-sample, two insights were obtained. First, in the individual setting, the round-number effects were only present for higher offer shares. Second, they manifested only for lower offer shares in the coordinative setting. Individual behavioral biases can easily explain the first insight, and the latter is apparently the result of coordinative considerations. The project contributes twofold to the literature. First, it provides an empirical analysis of millions of observations. Secondly, it describes a new experimental design that allows to study round-number effects.

In the third project, well–known experimental frameworks to identify the influence of gender and sex on economic decision–making were applied (Chapter 6). In a first step, data from participants whose gender and sex differ were collected on Prolific to study the impact of gender and sex on economic decisions. Then, the competitive, risk, and altruistic behavior of four different subject groups – cismen, ciswomen, transmen, and transwomen – was compared. In a next step, different priming conditions induced either a neutral, feminine, or masculine gender identity. Thus, with this experimental setup, the study went beyond correlating gender and sex with decisions and tried to understand causal relations through priming manipulation. It was concluded that the role of gender (and sex) is not as decisive for economic behavior as assumed. The results did not show conclusive correlational or causal evidence for gender or sex as determinants of economic decision–making. This project makes several contributions to the existing literature. First, it presents a sample of cis- and transgender participants in one framework covering their competitive, risk, and altruistic behavior. Second, the sample allows for both a correlational and a causal approach to study which part of gender effects can be attributed to biological factors (which refer to a participant's sex) and other aspects of one's gender identity. Finally, gender is measured not only on a categorical scale but also on a continuous gender scale that has been included in the analysis.

To draw a final conclusion to this dissertation, it remains to say that for the two main research areas - forecasting methods and behavioral economics - there are many more unexplored and unresolved questions. It would certainly be exciting and also tempting to pursue these questions. However, every journey must come to an end. The present work was a modest attempt to fit small pieces of the puzzle into the overall picture. I hope that I could convey my ideas and approaches convincingly to the reader, and thank you for your patience.

Appendices

Appendix A

Appendix: Chapter 3

A.1 Model architecture

These sections provide the formal description of the RNAA. The following notation follows Goodfellow et al. (2016) for one step i and the official `TensorFlow` documentation. `TensorFlow` operates on multi-dimensional arrays that are labeled tensors. It is possible to feed in only part of the data (batches) for computational efficiency. The batch size denotes the size of the parts, but is for readability set to 1 and omitted in this section. Activation functions are, in general, applied element-wise.

We assume the original definition of $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{x} = \left(x^{(1)}, \ldots, x^{(n_x)}\right),$$
$$\mathbf{y} = \left(y^{(1)}, \ldots, y^{(n_y)}\right),$$

where n_x and n_y denote the sequence lengths, which might vary across periods. The arguments of the activation functions, e.g., tanh, sigmoid, are regulated by a set of weights and biases. To provide a systematic naming concept that shows which input a weight matrix transfers to the activation function, the term *map* is a proxy for the term activation function. Therefore, in this section, b denotes a bias vector, U an input-to-map weight matrix, W an output-to-map weight matrix, V a state-to-map weight matrix and C a context-to-map weight matrix. The superscripts will distinguish the different weight matrices.

A.1.1 Encoder (LSTM)

The Encoder cell's architecture is an LSTM and the Encoder consists of multiple units representing the dimension of the states. The LSTM cell updates its cell state by the fraction of the previous state, $s^{(i-1)}$, determined by the *forget gate*, $f^{(i)}$, and the fraction of a proposed candidate cell state, $\tilde{s}$, determined by the *input gate*, $g^{(i)}$:

$$s^{(i)} = f^{(i)} \circ s^{(i-1)} + g^{(i)} \circ \tilde{s}^{(i)},$$

where $\circ$ denotes the element-wise multiplication and the candidate cell state is given by

$$\tilde{s}^{(i)} = \sigma\left(b^e + U^e \cdot x^{(i)} + W^e \cdot h^{(i-1)}\right),$$

where $\sigma(\cdot)$ denotes as usual a logistic sigmoid function.

The forget gate is computed with the input, $x^{(i)}$, and the previous output, $h^{(i-1)}$, of the cell for each step by

$$f^{(i)} = \sigma\left(b^f + U^f \cdot x^{(i)} + W^f \cdot h^{(i-1)}\right),$$

and the input gate uses a similar functional form but with its own weight matrices:

$$g^{(i)} = \sigma\left(b^g + U^g \cdot x^{(i)} + W^g \cdot h^{(i-1)}\right).$$

The output of the LSTM cell is computed by the hyperbolic tangent, $\tanh(\cdot)$, of the state by

$$h^{(i)} = \tanh\left(s^{(i)}\right) \circ q^{(i)},$$

and controlled by the *output gate*,

$$q^{(i)} = \sigma\left(b^o + U^o \cdot x^{(i)} + W^o \cdot h^{(i-1)}\right).$$

The cell state is initialized by $s^{(0)} = 0$. The dimensions of the matrices and vectors are as follows. Let m be the number of features, then

$$x^{(i)} : m \times 1.$$

Let n_e be the number of units of the Encoder, then

$$\begin{aligned} b^e, b^f, b^g, b^o &: n_e \times 1, \\ U^e, U^f, U^g, U^o &: n_e \times m, \\ W^e, W^f, W^g, W^o &: n_e \times n_e. \end{aligned}$$

The number of units, n_e is often referred to as *latent dimension*. So, the total number of parameters of the LSTM layer can be computed with

$$n_{\text{Encoder}} = 4(n_e + n_e m + n_e^2). \tag{A.1}$$

A.1.2 Decoder (GRU)

The main difference between the GRU and LSTM cell is the missing output gate in the former. Hence, the state and the output are the same. These two terms, state and output, are often used interchangeably in the literature, even though it is only correct for the GRU architecture. The GRU cell updates its current state by the fraction of the previous state and a candidate state by the *update gate*

$$s_d^{(i)} = \left(1 - z^{(i)}\right) \circ s_d^{(i-1)} + z^{(i)} \circ \tilde{s}_d^{(i)},$$

where the initial state, $s_d^{(0)}$, is the finale state of the Encoder. The candidate state, $\tilde{s}_d^{(i)}$, is provided as weighted sum of the previous target sequence step, the previous state weighted by the *reset gate* and the context vector,

$$\tilde{s}_d^{(i)} = \tanh\left(U^p \cdot y^{(i-1)} + V^p \cdot \left[r^{(i)} \circ s_d^{(i-1)}\right] + C^p \cdot c^{(i)} + b^p\right).$$

Notice that by default the GRU utilizes the ground truth, $y^{(i-1)}$, which is also included additionally to the previous state/output. The *update gate* is given by

$$z^{(i)} = \sigma\left(U^z \cdot y^{(i-1)} + V^z \cdot s_d^{(i-1)} + C^z \cdot c^{(i)} + b^z\right)$$

and the *reset gate* by

$$r^{(i)} = \sigma\left(U^r \cdot y^{(i-1)} + V^r \cdot s_d^{(i-1)} + C^r \cdot c^{(i)} + b^r\right).$$

In addition to the classical GRU structure, these gates contain the context vector, $c^{(i)}$, with the context-to-map weight matrix C. The context vector is provided by the attention mechanism as sum of the complete sequence of outputs of the Encoder weighted by the attention weights, $\alpha_j^{(i)}$. First, the sequence is processed by the Encoder and its output $\mathbf{h} = \left(h^{(1)}, \ldots, h^{(n_x)}\right)$ is obtained. Bahdanau et al. (2015) label $\mathbf{h}$ as *annotations*. Then, the context is computed with

$$c^{(i)} = \sum_{j=1}^{n_x} \alpha_j^{(i)} \cdot h^{(j)}$$

where the attention weights, $\alpha_j^{(i)}$, are computed using a softmax activation function,

$$\alpha_j^{(i)} = \frac{\exp\left(e_j^{(i)}\right)}{\sum_{k=1}^{n_x} \exp\left(e_k^{(i)}\right)}$$

based on the previous state of the Decoder, $s_d^{(i-1)}$, and all outputs of the Encoder

$$e_j^{(i)} = v_a' \cdot \tanh\left(V^a \cdot s_d^{(i-1)} + W^a \cdot h^{(j)} + b^a\right).$$

The output of the Decoder is computed by

$$h_d^{(i)} = \text{linear}\left(V^d \cdot s_d^{(i)} + C^d \cdot c^{(i)} + b^d\right),$$

where $linear(\cdot)$ returns its argument unchanged. The function is labeled for easier adjustment, e.g., to be replaced by the wide used softmax function.

Let n_d be the number of units in the Decoder, and n_a the number of units of the attention mechanism, and $y^{(i)}$ be a $m \times 1$ vector, then

$$\begin{aligned}
V^d &: m \times n_d,\\
C^d &: m \times n_e,\\
b^d &: m \times 1,\\
b^p, b^z, b^r &: n_d \times 1,\\
V^p, V^z, V^r &: n_d \times n_d,\\
U^p, U^z, U^r &: n_d \times m,\\
C^p, C^z, C^r &: n_d \times n_e,\\
b^a, v^a &: n_a \times 1,
\end{aligned}$$

$$V^a : n_a \times n_d,$$
$$W^a : n_a \times n_e.$$

The total number of parameters of the Decoder is

$$n_{\text{Decoder}} = m(4n_d + n_e + 1) + 3n_d(1 + n_d + n_e) + n_a(2 + n_d + n_e). \tag{A.2}$$

A.1.3 Number of parameters

The total number of parameters with normalization $n = n_a = n_e = n_d$ for the Encoder and Decoder is given by

$$n_{\text{Encoder}} = 4(n + nm + n^2), \tag{A.3}$$
$$n_{\text{Decoder}} = 5mn + m + 5n + 8n^2. \tag{A.4}$$

A.1.4 Deep learning layers

In the hyperparameter optimization, the Tuners are allowed to add additional layers after the Decoder to implement a very simple approach to a deep learning structure. This allows to search in a vertical and horizontal manner for an appropriate model. Each layer implements

$$h_l^{(i)} = \text{linear}\left(W^l \cdot h_{l-1}^{(i)} + b^l\right), \tag{A.5}$$

where $l = 1, \ldots, n_l$ denotes the additional dense layer. The output of the previous layer, $h_l^{(i)}$ is for $l = 0$ the output of the Decoder, $h_d^{(i)}$, and for each next layer, the previous deep learning layer, $h_l^{(i)}$, $l > 0$. The dimension of W^l differs with its position in the model. The first weight matrix for $l = 1$ has dimension $n_p \times m$ to transfer the output to the n_p cells of each dense layer. The last matrix for $l = n_l$ has dimension $m \times n_p$ to transfer to the number of features of the target sequence. The possible layers within the deep learning part are standardized to $n_p \times n_p$. The bias term, b^l adjusts accordingly to match the dimensions of W^l.

A.2 Additional data visualization

This appendix section provides more details on the primary data set.

A.2.1 Descriptives of the data set

Table A.1 summarizes the complete data set by the auction format, the product and the period for which the product is offered. It also provides the weekly or daily averages of the bid, i.e., the capacity and energy price as well as the bid size in MW.

Table A.1. Descriptive statistics of the data set.

					Averages		
Auction format	Product	Slice	Time	n	Capacity price	Energy price	MW offer
Weekly	NEG	HT	08:00 - 20:00	368	257.36	1249.39	18.07
		NT	20:00 - 08:00	368	500.10	1680.72	18.16
	POS	HT	08:00 - 20:00	368	293.20	1960.25	22.32
		NT	20:00 - 08:00	368	506.14	2009.66	21.48
Daily	NEG	-	00:00 - 04:00	845	16.19	738.39	6.72
		-	04:00 - 08:00	845	11.55	707.34	6.79
		-	08:00 - 12:00	845	7.98	537.85	6.98
		-	12:00 - 16:00	845	12.73	664.69	6.87
		-	16:00 - 20:00	845	7.28	515.27	7.06
		-	20:00 - 24:00	845	6.12	441.76	6.94
	POS	-	00:00 - 04:00	845	6.58	729.52	6.69
		-	04:00 - 08:00	845	9.25	828.81	6.66
		-	08:00 - 12:00	845	13.61	878.49	6.55
		-	12:00 - 16:00	845	10.07	714.61	6.66
		-	16:00 - 20:00	845	19.07	927.01	6.57
		-	20:00 - 24:00	845	11.73	860.17	6.64

Note: For each category, the number of observations, n, the average capacity price, energy price, and MW offer is reported.

A.2.2 MW capacity

Fig. A.1 shows the empirical cumulative distribution (ECDF) for the two auction formats for POS HT and POS 8:00 - 12:00. It can be seen that in both formats there is variation in the bid sizes. In the daily format, the 5 MW bid size is mostly observed.

A.2.3 Average weighted weekly capacity prices

Fig. A.2 shows the AWWC price for the POS HT of weekly auctions. The time series shows huge volatility and a downwards trend.

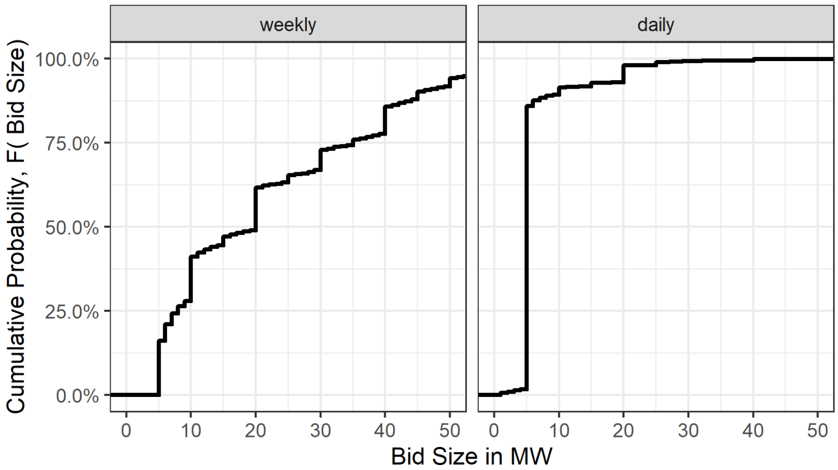


Figure A.1. Empirical cumulative distribution function of the bid size by auction format.

Note: Both plots show the range from 0 to 50, which covers 94.3% and 99.9% of the observations for the weekly and daily format, respectively.

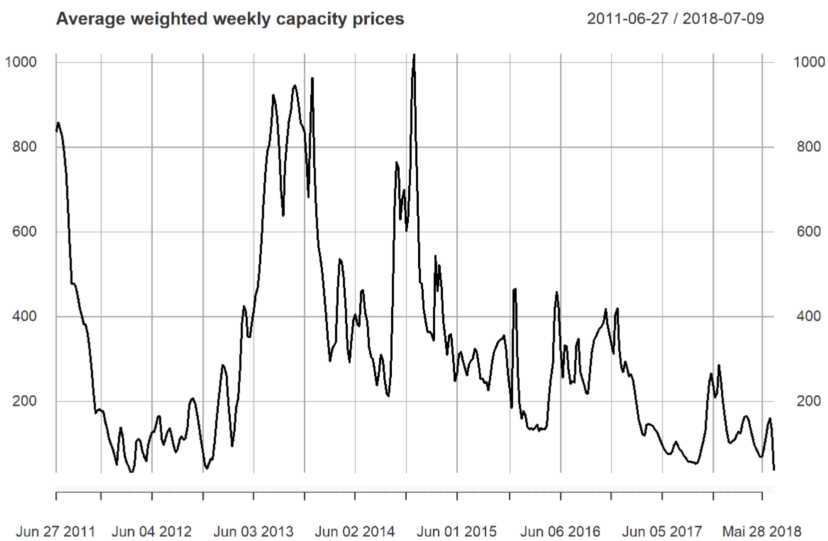


Figure A.2. Average weighted weekly capacity prices.

Note: The prices in €/MW according to Eq. (3.16) for the period from 2011/06/27 to 2018/07/09.

A.2.4 Varying sequence length

This section illustrates the problem that the sequence length is varying across time. It presents results based on weekly and daily auction formats. Fig. A.3 shows that there was an upwards trend in the number of bids. An interrupted time series analysis approach (red line) is applied on temporally equidistant observations, i.e., in the daily auction format, observations from one day per week are used. It identifies a significant positive linear trend and a significant positive shift of 168 bids when the auction format change was introduced. The interrupted time series approach (McDowall et al., 2019) shows a significant positive linear trend for the number of bids with a stark increase of 168 bids for the change of the auction format from weekly to daily. A frequency plot of these sequence lengths, Fig. A.4, shows a sharp distinction between the observed lengths by the auction format.

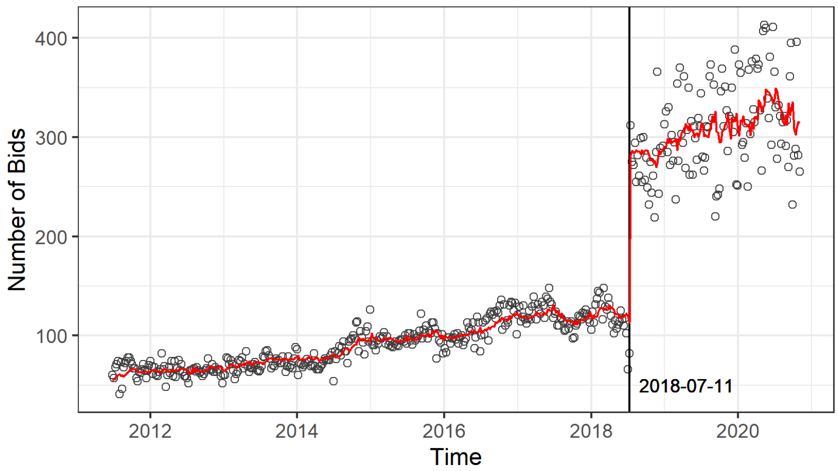


Figure A.3. Number of accepted bids.

Note: The number of bids of one supply curve represents the sequence length illustrated by gray circles. The solid black line marks the change from the weekly to the daily format on 2018/07/11. The red line shows the estimates of an interrupted time series analysis.

A.2.5 Supply meets demand

Fig. A.5 illustrates supply and demand. The total supply depicted by empty circles is the sum of each bid's capacity within one supply curve in one period. Red dots mark the GCC's announced demand, which was initially disclosed quarterly but changed to an a-day-before announcement which is not illustrated. Blue dots show the observed and published demand. The deviations in the first sections result from large capacity offers at the end of the supply curve with the highest capacity price from which only the required amount was procured. The GCC's change of the announcement method led to considerable variations in the last section, in which the exchange with Austria compensates for supply deficits or oversupply.

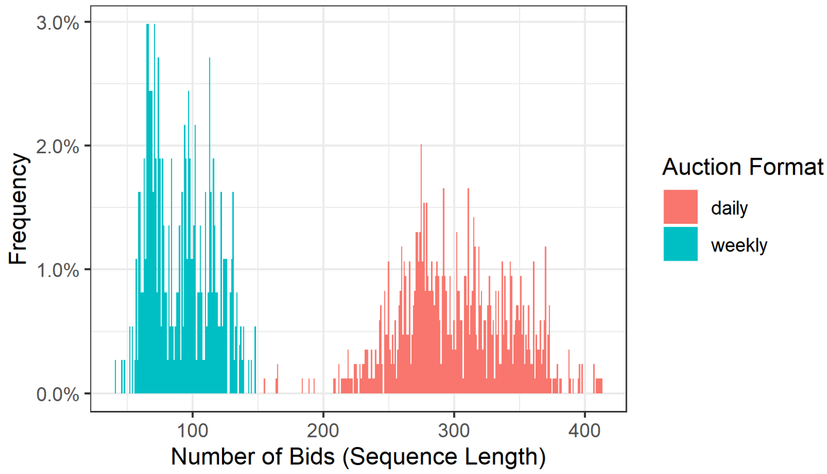


Figure A.4. Frequency plot of the number of accepted bids.

Note: The coloring indicates the auction format (red for daily, blue for weekly). There is a sharp distinction between the two formats separating in the left (weekly) and right (daily) concentrations. The average number of accepted bids is 91.05 with a standard deviation of 23.77 for the weekly auction and 300.8 with a standard deviation of 43.0 for the daily format.

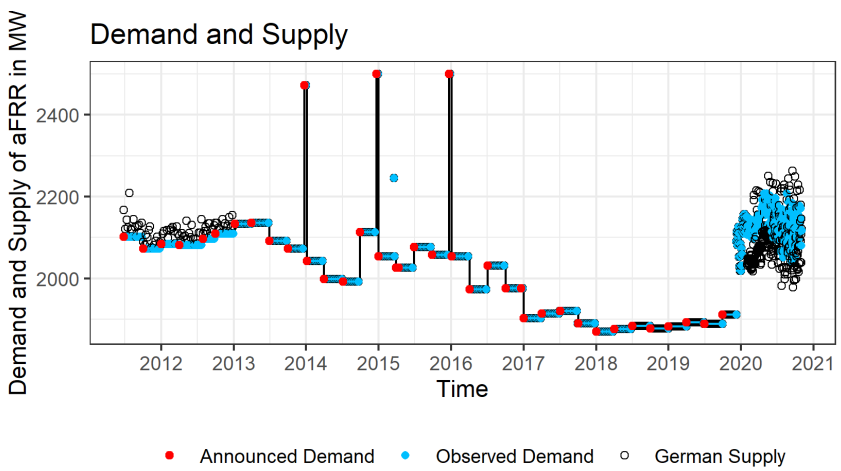


Figure A.5. Total supply, quarterly announced demand, and observed demand.

Note: Total supply, quarterly announced demand, and observed demand. Empty circles depict the total supply. Red dots mark the GCC's announced demand. Blue dots show the observed and published demand.

A.3 Additional results

A.3.1 Supply curve case with simple NNs

In the additional step of the performance evaluation, the aim is to predict the non-smoothed and non-truncated supply curves, including the bid sizes. In `Python`, this can be thought of as extending the last dimension of the 3D array by one more feature, or in other words, one supply curve is modeled as a two-column matrix; see Eq. (3.19). In the functional case, the evaluation of the smoothed curves on a fixed and constant number of points can be thought of as normalization to an equally spaced grid, elegantly omitting the irregular spacing problem. The RNAA does not have to rely on this normalization since it can process the bids directly, as just explained. Consequently, the RNAA is only compared to the default simple NN models in this section since other methods require more complex or combined approaches. The supply curves are taken as they were observed, and each sequence is padded with trailing zeros to the maximum length, as regularly done in the machine learning literature. The prices are logarithmized after adding an offset of +1 and cumulative bids are min-max normalized. The reported forecast metrics are computed on a Rolling principle but on a one-step basis instead of the complete curve.

All models have access to the complete input sequence, and the results are shown in Table A.2. The choice of hyperparameters for the RNAA remains unchanged compared to the functional case in Section 3.6.2. The Naive approach is replaced by the Last NN, which simply repeats the last observed step of the input sequence. The simple NNs utilize a Window approach for the data, which is adjusted to two scenarios. In the first scenario, the simple models are designed to predict the whole target sequences in a one-shot manner since the RNAA predicts the complete sequence step-by-step by default. In this scenario, the metrics decrease with increasing model complexity, as shown by the numbers in Table A.2 that become smaller line by line. Overall, the RNAA exhibits the smallest MSE and MAE compared to all benchmark models.

In a second scenario, it is accounted for the simplicity of the models, which might be the reason that they are unable to predict the whole sequence. So, the simple models are now designed to only predict a single step of the target sequence. This is very beneficial in terms of prediction errors since compared to the previous scenario, the MSE and MAE shown in Table A.2 are sharply reduced. However, the RNAA is still the model with the smallest MSE and outperforms three models in terms of MAE. Only the Last model has a smaller MAE, indicating fewer outliers, and is profiting the most from the reduced prediction task since only the last step is used naturally limiting deviations. Since the RNAA has access to the same input and consists of a step-by-step procedure, it is still comparable to these simple NNs, and it inherently addresses both scenarios simultaneously. So, the metrics for the RNAA remain the same.

As in the previous sections, the RNAA's attention mechanism produces an attention plot based on the training set. It is displayed in Fig. A.6. Compared to the attention plot

Table A.2. Forecast accuracy evaluation (supply curve case).

	One-shot		One-step	
Method	MSE	MAE	MSE	MAE
Last	0.22331	0.28263	0.09362	0.04918
Linear	0.11433	0.21807	0.07450	0.14197
Conv	0.10411	0.19799	0.05562	0.09038
LSTM	0.07538	0.15940	0.05562	0.06842
RNAA	0.04716	0.06537	0.04716	0.06537

Note: The table summarizes the MSE and MAE between the true and the predicted bids under the Rolling principle. The predictions are based on the simple NNs (Last, Dense, CONV, LSTM) and the RNAA in a one-shot and a one-step scenario. The Last method replaces the Naive approach but is essentially identical. The simple NNs are using a Window approach implementation. The reported metrics for the RNAA are identical in both scenarios because the RNAA constantly operates in a one-shot mode.

in the functional case, this plot shows that the middle section of the input sequences is very important for the final steps of the target sequence. Moreover, the deviations are visibly higher than in the functional case. Here, each step is a bid that consists of the capacity price and the capacity. So, the bids of the middle segment show high relevance when predicting the top bids of tomorrow and the starting bids. This is an intriguing insight, signaling that when using the supply curve only partly and restrict forecasts to the prices without considering the irregular spacing given by the bid sizes, valuable information is lost.

A.3.2 Evaluation to the same end

Table A.3 summarizes the MSABC for the supply curve case under the Rolling and Ahead forecasting principle approximated by Step and Linear. For this table, all methods are evaluated up to $\tau_L = 1$. So, it acts as a contrast to the approach presented in Section 3.6.3. It is hardly surprising that almost all MSABC for the Naive method and the RNAA decrease or remain identical. Solely for the Naive method, the MSABC under the Ahead principle with the Step approximation increases. It might have benefited more from the increased number of bids in the now truncated range.

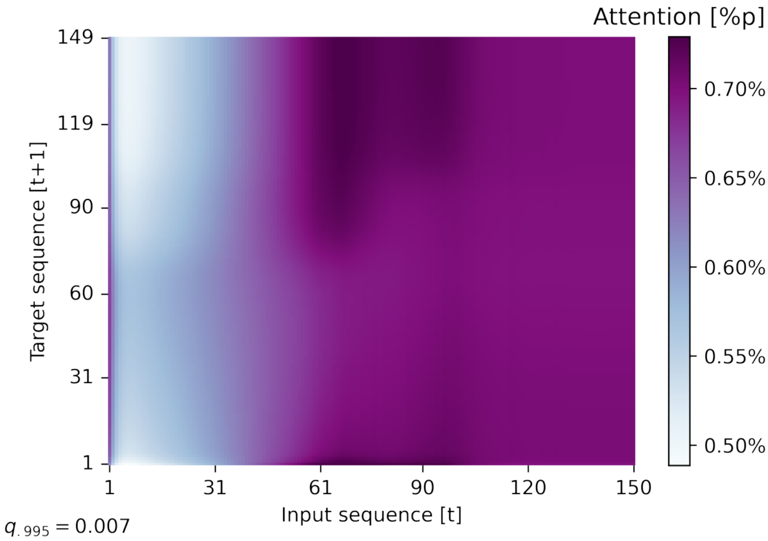


Figure A.6. Attention plot (supply curve case).

Note: The sequences consist of the non-smoothed supply curves, including the bid sizes and zero padding. The attention weights displayed in the attention plot are the average of all input and target training sequences across a bid's price and capacity. The model was trained on the data from 2011/06/27 to 2017/12/11. The coloring illustrated from light to dark shades the amount of attention assigned in ascending order. Attention weights over the 99.5% quantile are grouped in the darkest category for visualization reasons.

Table A.3. Forecast accuracy evaluation (supply curve case) with $\tau_L = 1$.

		Rolling		Ahead	
Range	Method	Step	Linear	Step	Linear
Total	Naive	0.00182	0.00190	0.00955	0.00910
	Bosq	0.00219	0.00207	0.01179	0.01117
	FTSA - Uni	0.00185	0.00171	0.00956	0.00904
	FTSA - Multi	0.00181	0.00168	0.00815	0.00770
	RNAA	0.00146	0.00140	0.00700	0.00658
Upper quartile	Naive	0.00010	0.00006	0.00063	0.00035
	Bosq	0.00020	0.00008	0.00077	0.00056
	FTSA - Uni	0.00018	0.00007	0.00063	0.00044
	FTSA - Multi	0.00016	0.00006	0.00057	0.00038
	RNAA	0.00019	0.00006	0.00017	0.00023

Note: The table summarizes the MSABC between the true curve and the predicted curve under the Rolling and Ahead principle. The FTSA methods (Bosq, FTSA - Uni, FTSA - Multi) and RNAA are evaluated up to 1 [truncated supply curve]. The column *Range* defines the starting point, i.e., the total range is from $\tau_0 = 0$ to $\tau_L = 1$. The upper quartile starts at the average upper quartile of 0.7598. The area under the curves is approximated by a stepwise connection of two points (Step) or by a linear interpolation (Linear).

A.4 Window approach

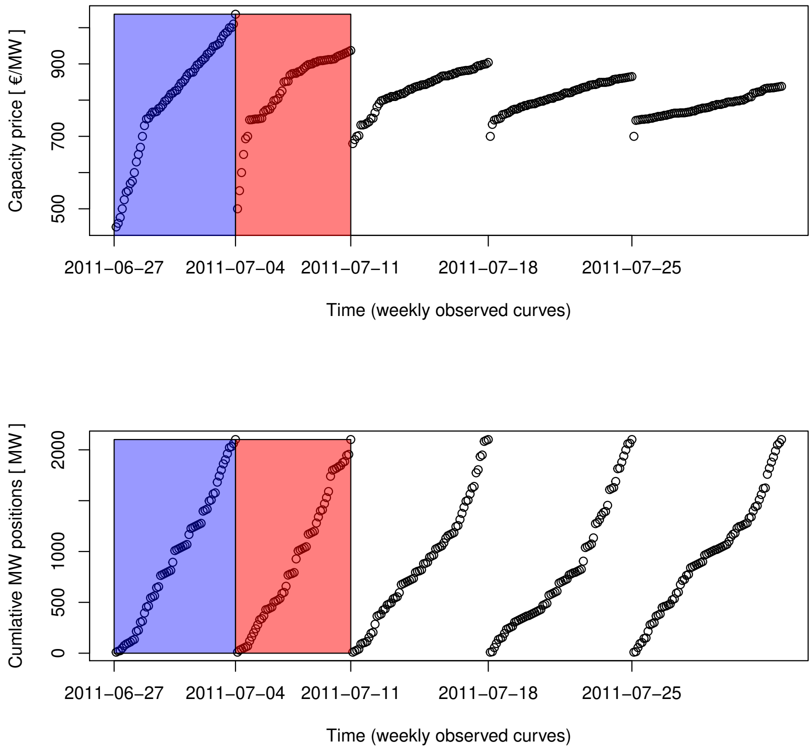


Figure A.7. Window approach on two-dimensional time series.

Note: The picture shows the first five observed supply curves. The supply curves are stacked in time, so the x-axis represents the time, where the intervals between two dates are subdivided by the bid numbers. The bid number is the position within a supply curve given by the increasing order of the capacity price. The top panel shows the capacity price dimension and the bottom panel the cumulative MW positions. The coloring illustrates the Window approach where the first curve is captured by the first, blue rectangle and the second, red rectangle captures the next curve. The former is the input and the latter is the target, which the network aims to predict. Both rectangles are shifted simultaneously by the same step size ahead along the time axis.

Appendix B

Appendix: Chapter 4

B.1 eBay data processing

Backus et al. (2020) provide two data sets **threads** and **lists**.[32] The term *thread* identifies a sequence of offers for a given buyer and seller pair bargaining over a given item. Each *observation* in the first data set consists of an offer from the buyer or the seller in a given thread with a time-stamp of the creation time and the response in plain text. Additionally, the current status of the offer (accepted, declined, countered) can be found. Hence, one thread can consist of multiple observations, but the last (most recent) observation covers the final decision and the final price, for which the bargaining parties settled. The second data set includes additional information about the item listing on the eBay platform. The unique item ID connects both data sets. The author's codebook provides more details on the covered variables.

We develop two filtering procedures that collect the data for the empirical analysis. The first one is written for the package `data.table` for `R` 3.6 and uses the data set threads. It creates a unique identifier for successful threads and calculates the duration of the negotiation, i.e., the time between the buyer's first offer and the final acceptance marking a successful thread. Additionally, it indicates whether the final price is round or not. The pseudo-code is shown in Algorithm 1. In particular, it first selects only the successful negotiations and then computes the duration for the different cases. Lastly, it creates the indicator for round numbers. The second algorithm is written for `STATA MP` 16.0, and the pseudo-code can be seen in Algorithm 2. It links the information of the item in the initial listing from the data set **lists** by using the unique item ID to the data set that the previous algorithm created.

In a last step, the complete data set is saved in the distribution-friendly csv-format and for the descriptive and regression analysis in the dta-format.

Table B.1 and Table B.2 summarize the distribution of the observations.

[32] The data set is publicly available at `Link` or Online (2022b).

Data: eBay data **threads**
Result: data set $(i, p_i, \Delta t_i, I(\cdot)_i))$
________________________;
for *all observations* ($n = 47,377,200$) **do**
    if *offer was accepted (status_id=1) or auto-accepted (status_id=9)* **then**
        keep ID of item, seller, buyer, thread;
        create new unique ID (i's) for each kept quadruple ($n = 12,018,417$)
    end
end
keep observations in **threads** by above new ID ($n = 17,892,293$);
transform plain text dates to interpretable dates;
for *all i with only one observation* **do**
    calculate Δt_i between creation date and response date;
    save by i: item id, price, Δt_i;
end
collect ($n = 8,534,338$);
for *all i with more than one observation* **do**
    order by creation date in ascending order;
    calculate Δt_i between creation date of first observation and response date of last observation;
    save by i: item id, price p_i, Δt_i;
end
collect ($n = 3,317,934$);
merge cases and reduce to one observation per ID ($n = 11,301,474$);
create $I(p_i \in \Upsilon)_i$;
remove duplicates($n = 4,159$);
save data set to external file;
get item's IDs ($n = 11,297,315$) and sort;
save item ID to external file;

Algorithm 1: Procedure in R 3.6.1.

```
Data: eBay data lists
Result: data set (i, X_i)
————————————————;
for all observations (n = 98,307,281) do
    order by item ID;
    if item ID is in item ID file then
        keep
    end
end
collect (n = 11,297,315);
merge with duration file (1:1);
```

Algorithm 2: Procedure in STATA 16.

Table B.1. Conditions of the items in the eBay data.

	N	%	$\sum$%
New	2,300,284	28.24	28.24
New other	695,968	8.55	36.79
New with defects	35,918	0.44	37.23
Manufacturer refurbished	12,032	0.15	37.38
Seller refurbished	36,985	0.45	37.83
Like New	315,149	3.87	41.70
Used	4,286,288	52.63	94.33
Very Good	219,391	2.69	97.02
Good	114,677	1.41	98.43
Acceptable	32,636	0.40	98.83
For parts / not working	95,047	1.17	100.00
Total	8,144,375	100.00	

Note: The table shows the distribution of the item's condition of the eBay data set of Backus et al. (2020) after applying our algorithm. The conditions are ordered by their numeric ID in the data set.

Table B.2. Categories of the items in the eBay data.

	N	%	$\sum$%
Collectibles	1,414,232	12.75	12.75
Everything else	42,679	0.38	13.14
Toys and Hobbies	658,335	5.94	19.07
Dolls and Bears	125,574	1.13	20.21
Stamps	106,335	0.96	21.16
Books	302,895	2.73	23.90
Jewelry and Watches	711,072	6.41	30.31
Consumer Electronics	263,319	2.37	32.68
Specialty Services	930	0.01	32.69
Art	108,317	0.98	33.67
Musical Instruments and Gear	178,088	1.61	35.27
Cameras and Photo	144,795	1.31	36.58
Pottery and Glass	209,958	1.89	38.47
Sporting Goods	421,476	3.80	42.27
Video Games and Consoles	184,018	1.66	43.93
Pet Supplies	14,453	0.13	44.06
Tickets and Experiences	28,727	0.26	44.32
Baby	28,417	0.26	44.58
Travel	10,717	0.10	44.67
Real Estate	81	0.00	44.67
Coins and Paper Money	283,656	2.56	47.23
DVDs and Movies	108,607	0.98	48.21
Music	212,624	1.92	50.13
Clothing Shoes and Accessories	2,487,553	22.43	72.56
Home and Garden	330,926	2.98	75.54
Business and Industrial	393,465	3.55	79.09
Crafts	93,174	0.84	79.93
Cell Phones and Accessories	135,474	1.22	81.15
Antiques	202,743	1.83	82.98
Health and Beauty	136,147	1.23	84.21
Entertainment Memorabilia	135,324	1.22	85.43
Computers or Tablets and Networking	356,458	3.21	88.64
Sports Mem Cards and Fan Shop	1,258,065	11.34	99.99
Gift Cards and Coupons	1,645	0.01	100.00
Total	11,090,279	100.00	

Note: The table shows the distribution of the item's category of the eBay data set of Backus et al. (2020) after applying our algorithm. The categories are ordered by their numeric ID in the data set.

B.1.1 Detailed eBay regression results

Table B.3. Detailed regression results of duration or number of periods on round prices.

	(1) Duration		(2) Duration		(3) Periods		(4) Periods	
Round numbers	-24.82***	(4.91)	-53.02***	(5.21)	-0.17***	(0.00054)	-0.19***	(0.00064)
Condition								
New			0.00	(.)			0.00	(.)
New other			11.65	(10.4)			-0.05***	(0.0014)
New with defects			199.56***	(45.3)			-0.01**	(0.0051)
Manufacturer refurbished			-34.89	(57.2)			0.09***	(0.0098)
Seller refurbished			-11.81	(33.6)			0.04***	(0.0054)
Like New			32.84	(21.0)			-0.05***	(0.0022)
Used			-79.84***	(5.98)			-0.10***	(0.00079)
Very Good			-40.32**	(20.5)			-0.10***	(0.0025)
Good			-51.99**	(26.2)			-0.11***	(0.0031)
Acceptable			-36.38	(43.4)			-0.11***	(0.0051)
For parts / not working			332.98***	(30.2)			-0.08***	(0.0033)
Category								
Collectibles			0.00	(.)			0.00	(.)
Everything else			-1159.55***	(9.80)			-0.59***	(0.0011)
Toys and Hobbies			-138.58***	(12.3)			0.10***	(0.0015)
Dolls and Bears			-310.26***	(20.5)			-0.04***	(0.0026)
Stamps			167.65	(394.9)			-0.06*	(0.037)
Books			-165.24***	(23.4)			-0.07***	(0.0025)
Jewelry and Watches			-18.92	(13.3)			0.13***	(0.0016)
Consumer Electronics			-191.30***	(14.6)			0.19***	(0.0022)
Art			81.57	(115.6)			0.02*	(0.0086)
Musical Instruments and Gear			-113.83***	(16.0)			0.22***	(0.0026)
Cameras and Photo			-162.89***	(18.5)			0.21***	(0.0029)
Sporting Goods			-194.64***	(12.6)			0.19***	(0.0018)
Video Games and Consoles			-374.33***	(17.3)			0.25***	(0.0028)
Pet Supplies			-246.58***	(54.7)			0.04***	(0.0078)
Baby			-363.80***	(20.7)			0.14***	(0.0058)
Travel			-185.42***	(42.2)			0.18***	(0.010)
Coins and Paper Money			30.11	(239.2)			0.01	(0.025)
DVDs and Movies			-324.70***	(23.1)			0.03***	(0.0031)
Music			8.51	(23.6)			-0.06***	(0.0020)
Clothing Shoes and Accessories			-206.50***	(9.79)			0.05***	(0.0010)
Home and Garden			-211.41***	(12.9)			0.06***	(0.0018)
Business and Industrial			530.35***	(21.2)			0.06***	(0.0017)
Crafts			-412.20***	(19.1)			-0.13***	(0.0027)
Cell Phones and Accessories			-645.60***	(11.2)			0.26***	(0.0030)
Antiques			120.52	(242.8)			-0.10***	(0.037)
Health and Beauty			-251.97***	(19.4)			0.01***	(0.0026)
Entertainment Memorabilia			255.81*	(137.1)			0.02**	(0.012)
Computers or Tablets and Networking			-256.83***	(12.6)			0.13***	(0.0019)
Sports Mem Cards and Fan Shop			149.11***	(18.7)			0.11***	(0.0020)
Constant	1059.31***	(3.02)	1165.92***	(9.80)	1.58***	(0.00038)	1.59***	(0.0011)
N	11,090,279		8,144,375		11,090,279		8,144,375	

Note: The table reports OLS results for the two dependent variables, Duration and Periods. *Duration* denotes the time between the first observation and the last observation of a thread in minutes. *Periods* denotes the number of offers made between seller and buyer. The table reports the coefficient of the round number dummy as *Round numbers*. There are 11 condition dummies for the item, where the baseline is "New". The meta category of the item has 38 categories and is considered with a corresponding number of dummies, where the baseline is "Collectible". Missing observations are due to incomplete recordings of condition or category. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

B.2 Additional experimental results

B.2.1 Acceptance and rejection times

In this section, we provide details on the decision times for acceptances and rejections separately in addition to Section 4.3. Table B.4 summarizes the decision times when an offer was accepted. The discussion can be found in Section 4.3.

Table B.4. Decision times conditional on acceptance.

		Treatment	
Offer type	Total	Single	Partner
Round	9.81	9.88	9.75
NonRound	11.05	10.55	11.52

Note. Average decision times are reported in seconds.

Table B.5 summarizes the decision times when an offer was rejected. When the observations of rejections are pooled across treatments, we find significantly quicker rejections when a round offer was made (t-test: 7.92s vs. 8.65s; $p = 0.0527$). Furthermore, the difference in decision times between offer types for rejections is the smallest (0.73s) compared to the previous cases when decisions were pooled (1.11s) or when only acceptances were considered (1.24s). When we control for the treatments, we find that participants in Partner reject round offers significantly quicker (t-test: 7.96s vs. 8.95s; $p = 0.0479$) while the difference in Single (t-test: 7.87s vs. 8.37s; $p = 0.3730$) is not significant.

Table B.5. Decision times conditional on rejection.

		Treatment	
Offer type	Total	Single	Partner
Round	7.92	7.87	7.96
NonRound	8.65	8.38	8.95

Note. Average decision times are reported in seconds.

B.2.2 Acceptance frequency bar plot

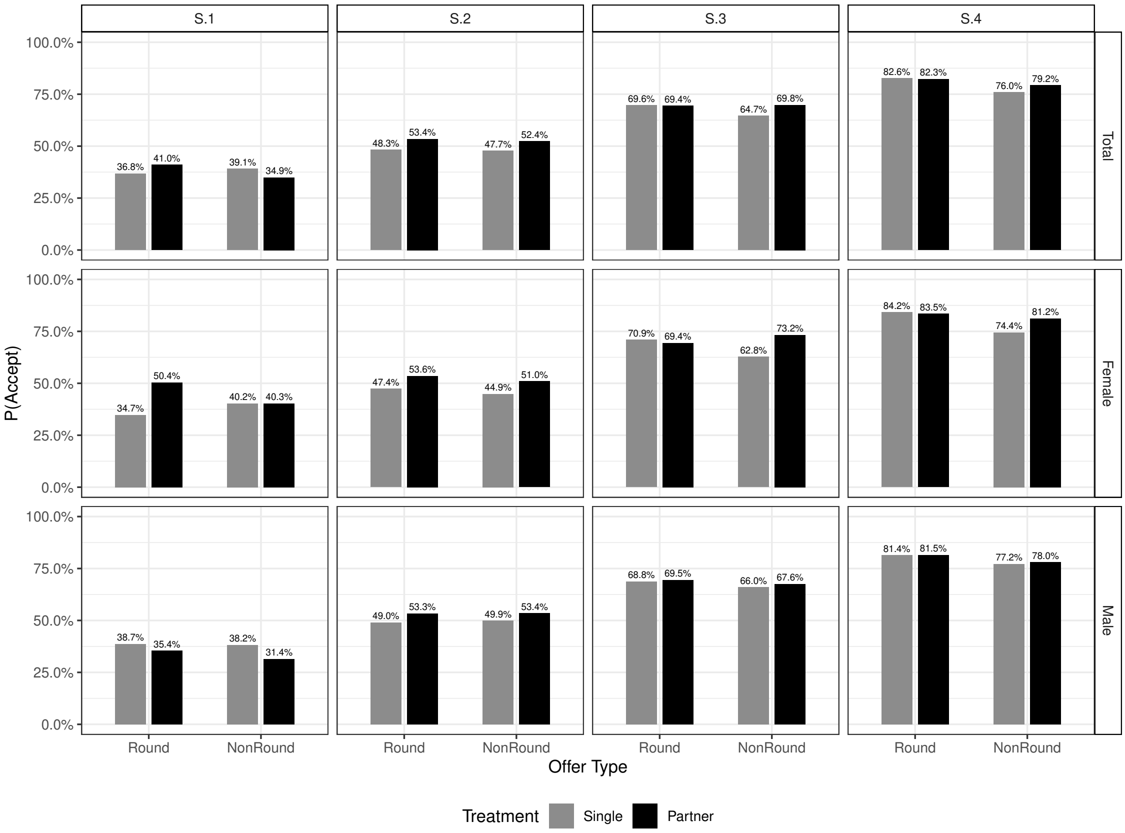


Figure B.1. Acceptance frequencies as bar plot for each segment.

Note: The rows of the figure correspond to the total, female and male sample as indicated by the right legend. Each column corresponds to a segment of offer shares (S.1, S.2, S.3, S.4). The segments are equally wide. For each cell of the figure, the share of accepted round and non-round offers in the two treatments is illustrated. The gray bars represent the Single treatment, and the black bars correspond to the Partner treatment.

B.2.3 Regression analysis

To evaluate the robustness of our results in Section 4.3, we estimate a linear probability model by OLS, where standard errors are clustered on the individual-level. The dependent variable is the binary variable offer acceptance. We control for offer share, treatment, offer type, and the interaction of the latter two.

As we have seen in our analysis from Fig. 4.7, a higher propensity to accept is associated with round numbers in all treatments, but they are likely caused by different channels for high and low offer shares. Thus, it is not surprising to see significant round-number effects and no significant interaction term without restricting the offer share, as the round-number dummy simply captures the whole round-number effect. To control for this and to keep the estimated models as parsimonious as possible, we divide our sample into the four offer share segments as before and estimate the same model separately for each segment. Table B.6 shows the results of the estimations. The first three columns summarize the estimates for

S.1. We find significant positive interaction terms. In the total sample, being in Partner and receiving a round offer increases the acceptance frequency by 8.4%p ($p = 0.080$) on average. The effect is especially pronounced in the female sample, where, ceteris paribus, a round offer has a 15.7%p higher chance of being accepted ($p = 0.035$). Thus, round numbers have a higher acceptance frequency in Partner for lower offer shares (**Result 2**). For S.2, we do not find any significant treatment or round-number effects. Again, this is in line with the graphical analysis (Column (4) to (6)). For S.3 (Column (7) to (9)), in the total sample, there are significant treatment and round number effects. The interaction is insignificant. This confirms our argument for round-number bias (**Result 1**). Only looking at the female sample, the results are qualitatively similar, but now the interaction is significantly negative. This could already be seen in the figures above and further confirms our conjecture that in Partner, subjects were more careful in their decision-making, thereby reducing potentially unconscious biases for round numbers. For S.4, we get qualitatively similar results. Only now the interaction term for the female sample becomes insignificant, yet still has the negative sign as in S.3 (Column (10) to (12)).

Table B.6. OLS Regression for segments. Dependent variable: Offer acceptance.

	Segment 1			Segment 2			Segment 3			Segment 4		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Sample:	Total	Female	Male	Total	Female	Male	Total	Female	Male	Total	Female	Male
(Intercept)	0.420 ***	0.354 ***	0.461 ***	0.129	0.069	0.172	0.311 ***	0.195	0.386 ***	0.606 ***	0.809 ***	0.469 ***
	(0.056)	(0.085)	(0.076)	(0.081)	(0.126)	(0.106)	(0.096)	(0.152)	(0.124)	(0.087)	(0.129)	(0.117)
Offer Share	-0.125	0.208	-0.337	0.804 ***	0.881 ***	0.754 ***	0.513 ***	0.661 ***	0.420 **	0.168 *	-0.071	0.332 ***
	(0.203)	(0.313)	(0.265)	(0.174)	(0.273)	(0.228)	(0.142)	(0.226)	(0.182)	(0.094)	(0.141)	(0.125)
Treatment: Partner	-0.042	-0.000	-0.068	0.046	0.061	0.035	0.051 *	0.105 **	0.016	0.033	0.068 **	0.008
	(0.041)	(0.064)	(0.052)	(0.033)	(0.050)	(0.043)	(0.029)	(0.044)	(0.039)	(0.020)	(0.031)	(0.026)
Round Offer	-0.027	-0.049	-0.005	0.007	0.019	-0.003	0.048 *	0.080 **	0.026	0.076 ***	0.094 **	0.061 *
	(0.033)	(0.049)	(0.046)	(0.033)	(0.049)	(0.046)	(0.025)	(0.039)	(0.032)	(0.025)	(0.037)	(0.034)
Treatment: Partner x Round Offer	0.084 *	0.157 **	0.035	0.005	0.014	0.000	-0.053	-0.118 **	-0.010	-0.036	-0.076	-0.009
	(0.048)	(0.074)	(0.063)	(0.046)	(0.071)	(0.062)	(0.036)	(0.058)	(0.047)	(0.033)	(0.050)	(0.044)
N	1520	646	874	2120	899	1221	2464	988	1476	3136	1287	1849
R2	0.003	0.013	0.006	0.013	0.016	0.011	0.008	0.017	0.004	0.005	0.009	0.005

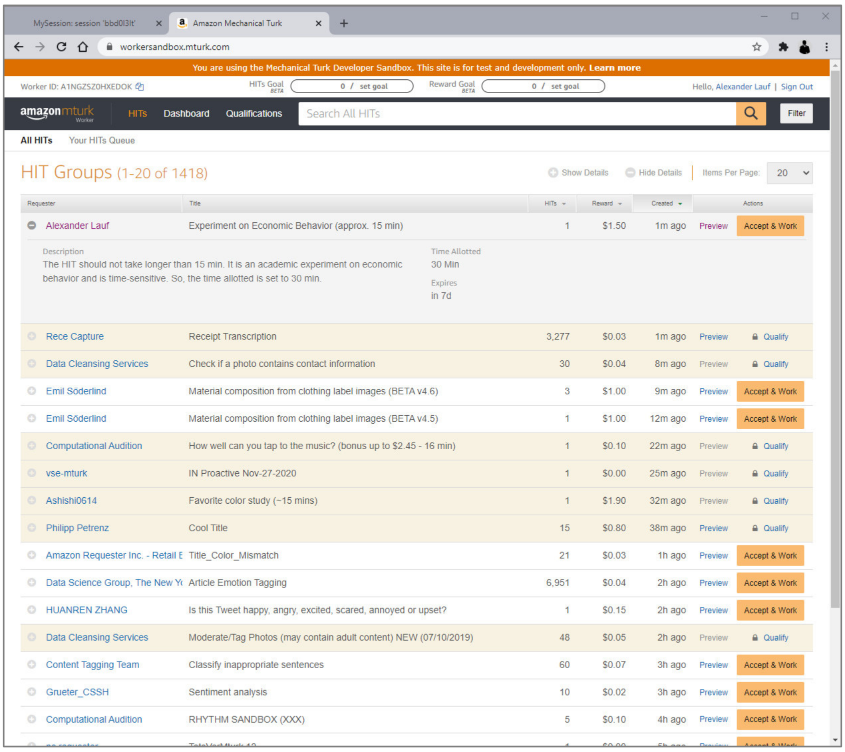
Note: The offer acceptance is a binary variable equal to 1 if the participant accepted an offer and 0 otherwise. Standard errors in parentheses are heteroskedasticity robust and clustered on the individual level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

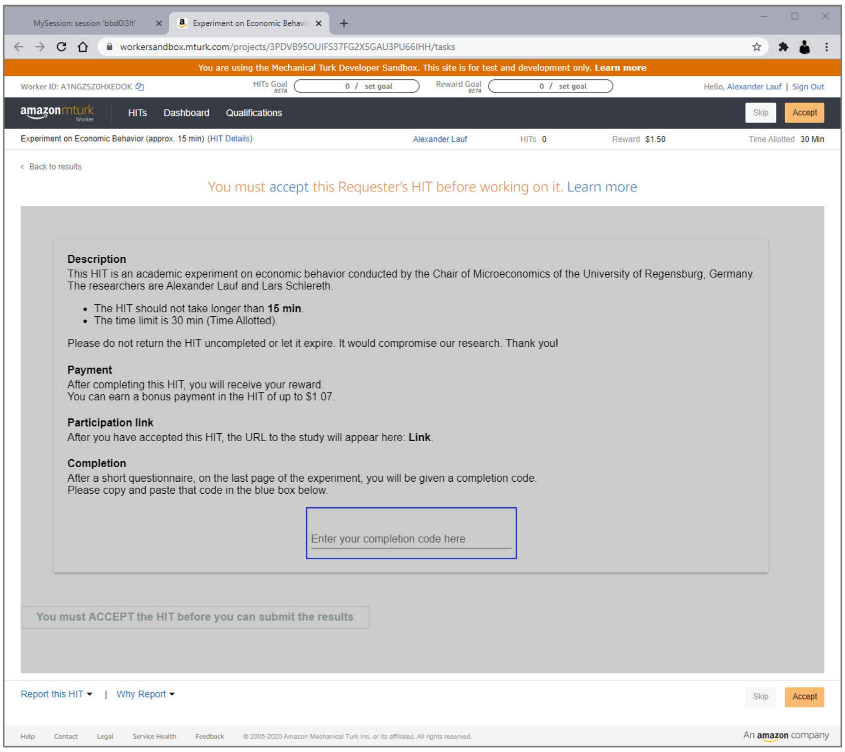
B.3 MTurk and oTree instructions

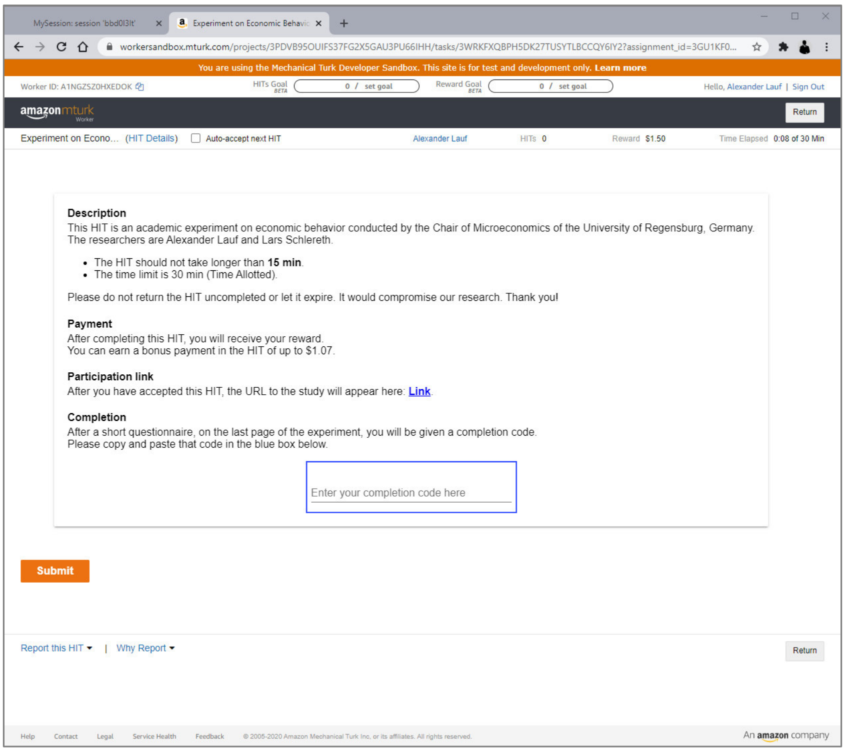
This section provides screenshots of the human intelligence task (HIT) published on Amazon Mechanical Turk under the name of Alexander Lauf as the requester and the instructions of the experiment in oTree for both treatments. Please note that these pictures represent websites. The oTree code is available on request. In particular, the following is covered:

1. HIT - Design and description,
2. Experimental design: Single,
3. Experimental design: Partner.

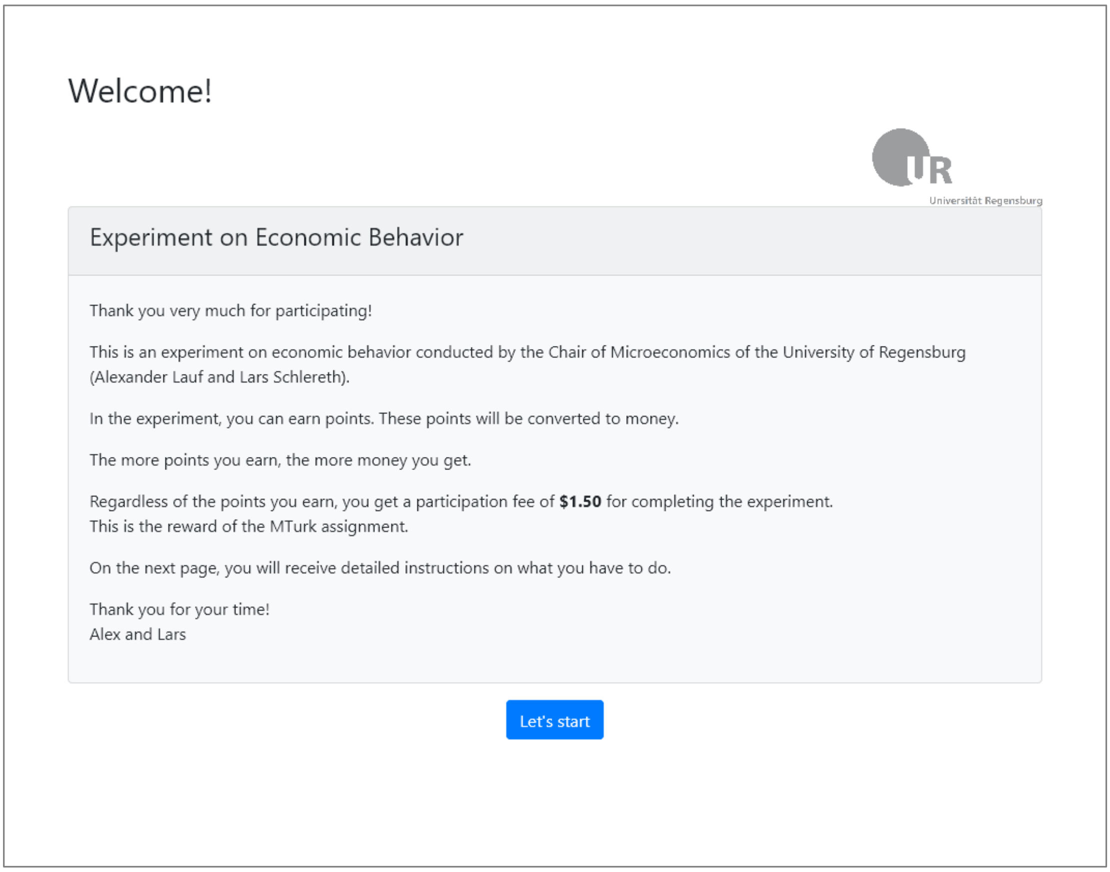
B.3.1 HIT - Design and description






Experiment on Economic Behavi
workersandbox.mturk.com/projects/3PDVB95OUIFS37FG2X5GAU3PU66IHH/tasks/3WRKFXQBPH5DK27TUSYTLBCCQY6IY2?assignment_id=3GU1KF0...
You are using the Mechanical Turk Developer Sandbox. This site is for test and development only. Learn more
Worker ID: A1NGZSZ0HXEDOK
HITs Goal
Reward Goal
Hello, Alexander Lauf | Sign Out
Return
Experiment on Econo... (HIT Details)
Auto-accept next HIT
Alexander Lauf
HITs 0
Reward $1.50
Time Elapsed 0:08 of 30 Min
Description
This HIT is an academic experiment on economic behavior conducted by the Chair of Microeconomics of the University of Regensburg, Germany. The researchers are Alexander Lauf and Lars Schlereth.
• The HIT should not take longer than 15 min.
• The time limit is 30 min (Time Allotted).
Please do not return the HIT uncompleted or let it expire. It would compromise our research. Thank you!
Payment
After completing this HIT, you will receive your reward.
You can earn a bonus payment in the HIT of up to $1.07.
Participation link
After you have accepted this HIT, the URL to the study will appear here: Link
Completion
After a short questionnaire, on the last page of the experiment, you will be given a completion code.
Please copy and paste that code in the blue box below.
Enter your completion code here
Submit
Report this HIT | Why Report
Return
Help Contact Legal Service Health Feedback
An amazon company

B.3.2 Experimental design: Single



Welcome!

Universität Regensburg

Experiment on Economic Behavior

Thank you very much for participating!

This is an experiment on economic behavior conducted by the Chair of Microeconomics of the University of Regensburg (Alexander Lauf and Lars Schlereth).

In the experiment, you can earn points. These points will be converted to money.

The more points you earn, the more money you get.

Regardless of the points you earn, you get a participation fee of **$1.50** for completing the experiment.
This is the reward of the MTurk assignment.

On the next page, you will receive detailed instructions on what you have to do.

Thank you for your time!
Alex and Lars

Let's start

Instructions

What do you have to do?

In this experiment, you will make one decision in each of **10 periods**, i.e., a total of 10 decisions. In this experiment, you can earn points. After this experiment, the points will be converted into real money **(1000 Points = 1 Dollar)** and paid to you as a bonus later.

The points you earn in the experiment depend on your own decisions and the decisions of the computer in a way explained below.

In each period, you will receive an **offer**. The computer **randomly selects** one of the numbers in a blue box below. Each number is **equally likely** to be drawn. You can see the box for Period 1 below.

You can **accept** or **reject** the offer.

Line 1	1067	1033	1000
Line 2	967	933	900
Line 3	867	833	800
Line 4	767	733	700
Line 5	667	633	600
Line 6	567	533	500
Line 7	467	433	400
Line 8	367	333	300
Line 9	267	233	200
Line 10	167	133	100

Period 1

In each next period, **numbers are removed** from period to period. In Period 2, 'Line 1' is removed. In Period 3, additionally, 'Line 2' is removed. This is repeated until period 10, in which only the last line remains.

In each period, you decide to either **accept** or **reject** the offer. In subsequent periods, you can get a higher or lower offer, but keep in mind that the **three highest numbers** are removed in each period.

You can click on the right-arrow in the slider gallery below to see how these lines are removed.

Line 1	~~1067~~	~~1033~~	~~1000~~
Line 2	967	933	900
Line 3	867	833	800
Line 4	767	733	700
Line 5	667	633	600
Line 6	567	533	500
Line 7	467	433	400
Line 8	367	333	300
Line 9	267	233	200
Line 10	167	133	100

Period 2

After you have made your decisions for all 10 periods, your bonus in points is determined as follows:

The computer **randomly selects** a period out of Period 1 to 10. Then, the computer moves step-by-step from **this randomly selected period** to Period 10. The computer stops in the period, where for the first time you have accepted the respective period's offer. This offer is your bonus in points.

Therefore, you should play every period as if the computer selected the current period.

Your total payment in dollars is the sum of participation fee and your bonus in points converted to dollars (1000 Points = 1 Dollar):

$$\text{Payment} = \text{Participation Fee} + \frac{\text{Your Points}}{1000}$$

Now we ask you a few check-up questions. They do not have any impact on your payment. After you have finished the experiment, we will conduct a short survey.

Go to Check-Up Questions

Check-Up Question | Instructions

Check-Up Questions

Question (1)

The following questions have no effect on your amount of points. Please read carefully.

We want to check whether you understood the instructions. If you want to read the instructions again, you can find a tab 'Instructions' in the navigation bar above.

Suppose you are in **Period 1**. You got your offer from the numbers in the blue boxes below. You are about to move to the next period.

Line 1	1067	1033	1000
Line 2	967	933	900
Line 3	867	833	800
Line 4	767	733	700
Line 5	667	633	600
Line 6	567	533	500
Line 7	467	433	400
Line 8	367	333	300
Line 9	267	233	200
Line 10	167	133	100

Period 1

Which line is eliminated when you move to the next round (Period 2)?

- ○ Line 1: 1067, 1033, 1000
- ○ Line 5: 667, 633, 600
- ○ Line 9: 267, 233, 200
- ○ A random Line.

Instructions | Next Question

Check-Up Question | Instructions

Check-Up Questions

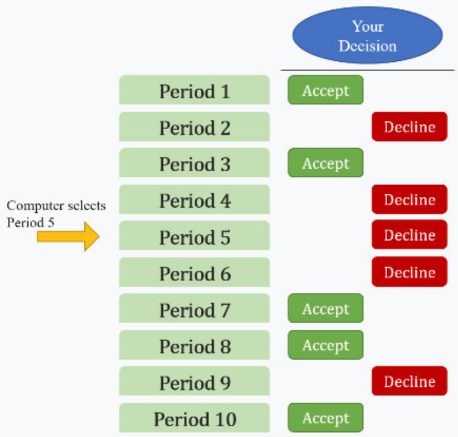
Question (2)

The following questions have no effect on your amount of points. Please read carefully.

We want to check whether you understood the instructions. If you want to read the instructions again, you can find a tab 'Instructions' in the navigation bar above.

Suppose you have already made **all decisions**. In the picture below, you can see all of your decisions.

The computer has selected **Period 5**. As mentioned in the instructions, the computer starts in Period 5 and goes step-by-step to Period 10.



From which period do you get the offer as final payment?

- ○ Period 5
- ○ Period 10
- ○ Period 7
- ○ A random Period.

Instructions | Next Question

Check-Up Question | Instructions

Check-Up Questions

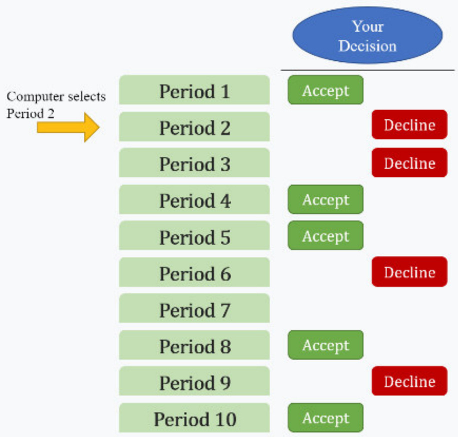
Question (3)

The following questions have no effect on your amount of points. Please read carefully.

We want to check whether you understood the instructions. If you want to read the instructions again, you can find a tab 'Instructions' in the navigation bar above.

Suppose you have already made **all decisions**. In the picture below, you can see all of your decisions.

The computer has selected **Period 2**. As mentioned in the instructions, the computer starts in Period 2 and goes step-by-step to Period 10.



From which period do you get the offer as final payment?

- ○ Period 4
- ○ Period 8
- ○ Period 9
- ○ A random Period.

Instructions | Next Question

Check-Up Question | Instructions

Check-Up Questions

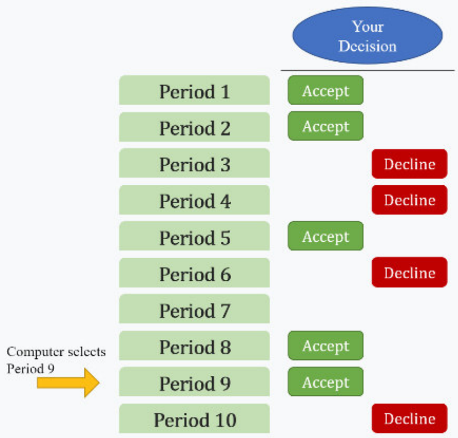
Question (4)

The following questions have no effect on your amount of points. Please read carefully.

We want to check whether you understood the instructions. If you want to read the instructions again, you can find a tab 'Instructions' in the navigation bar above.

Suppose you have already made **all decisions**. In the picture below, you can see all of your decisions.

The computer has selected **Period 9**. As mentioned in the instructions, the computer starts in Period 9 and goes step-by-step to Period 10.



From which period do you get the offer as final payment?

- ○ Period 2
- ○ Period 9
- ○ Period 5
- ○ A random Period.

Instructions | Continue

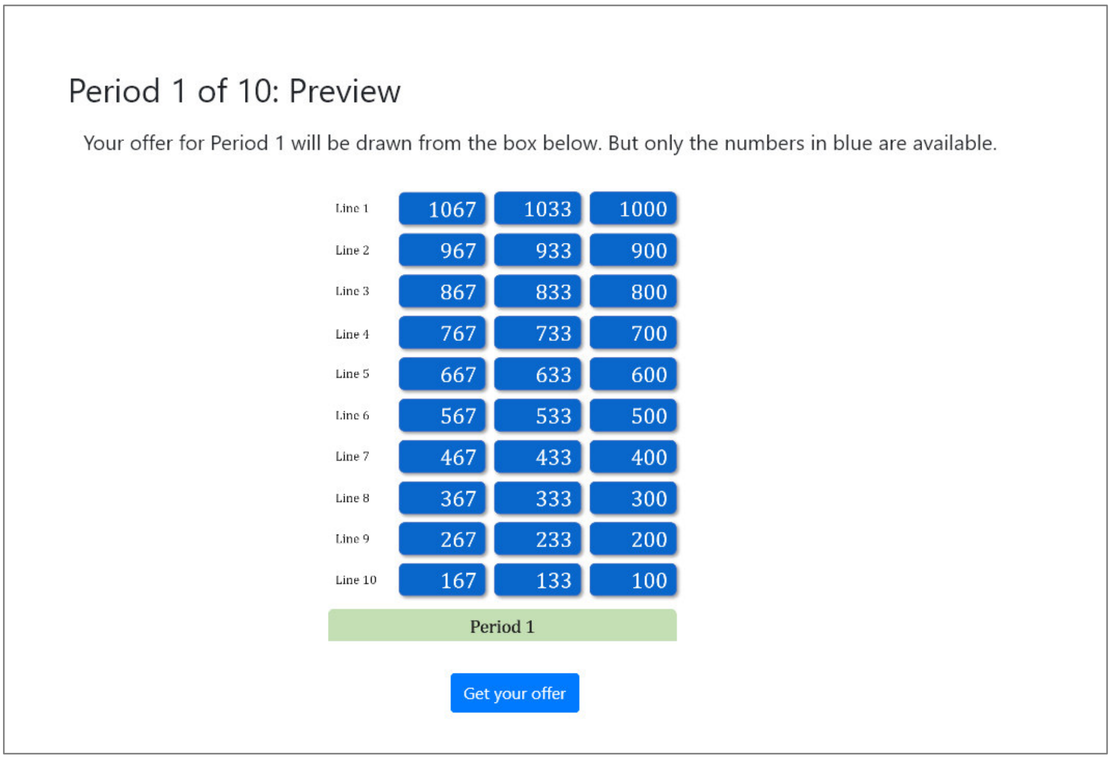
Start of the Experiment

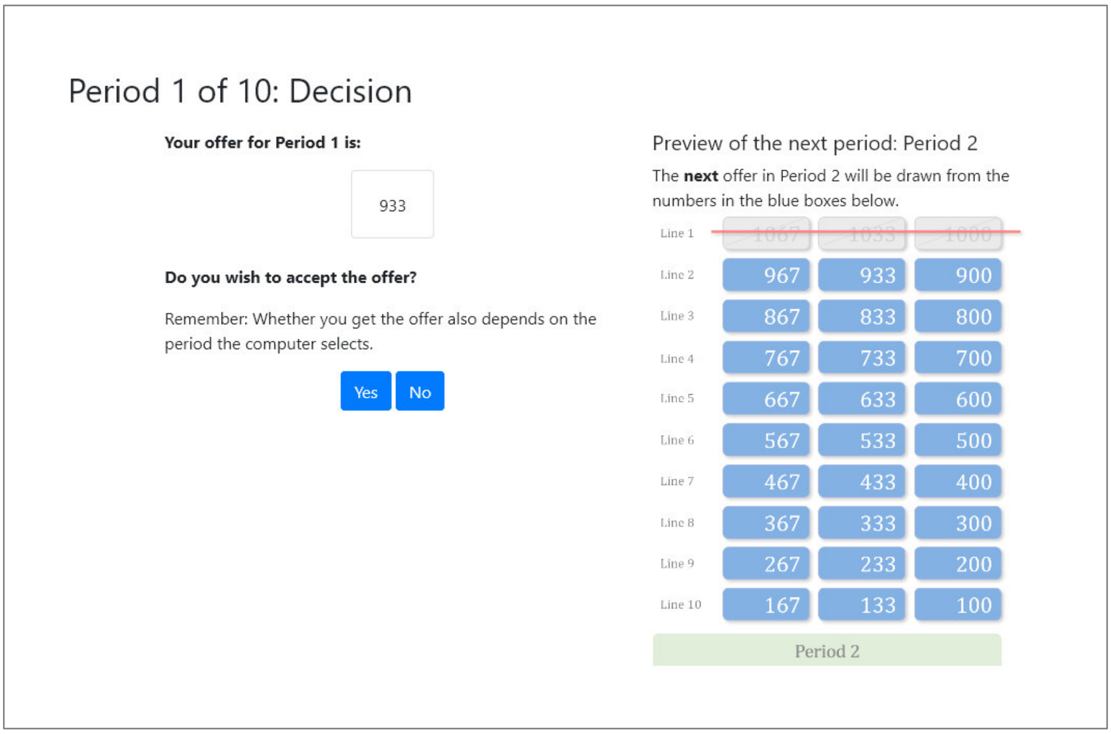
You have finished the check-up questions.

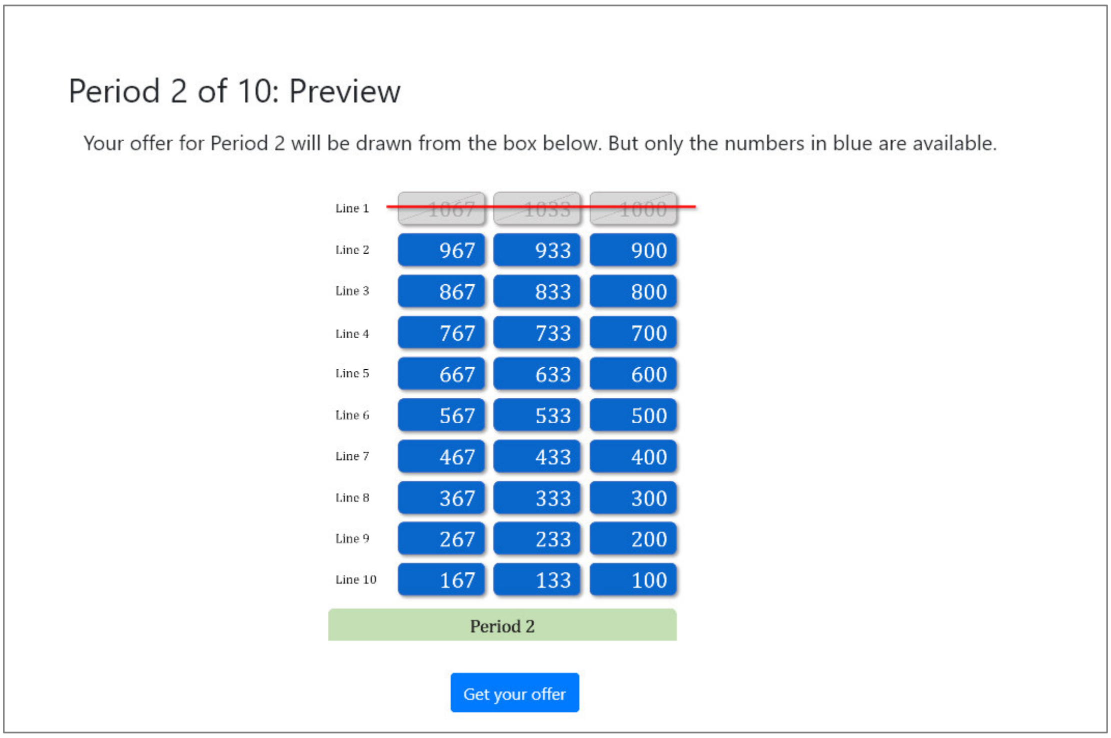
You can now start the experiment by clicking on the button below.

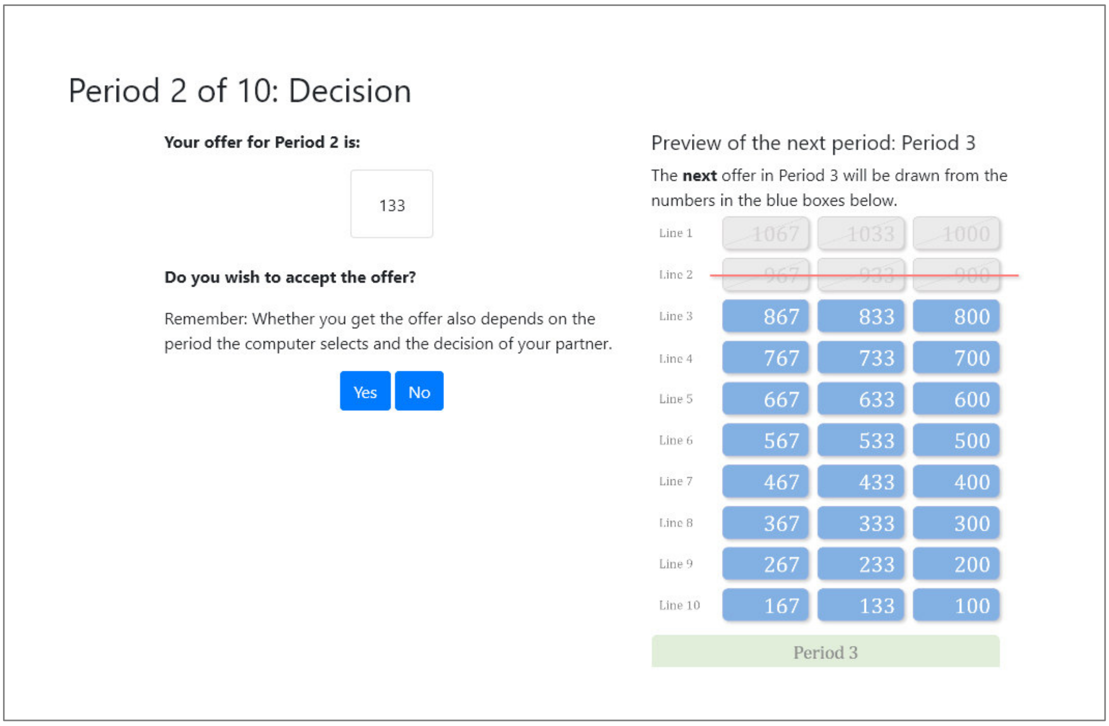
All your answers to the following questions will be relevant for your final payoff.

Start the Experiment!


Period 1 of 10: Preview
Your offer for Period 1 will be drawn from the box below. But only the numbers in blue are available.
Line 1 1067 1033 1000
Line 2 967 933 900
Line 3 867 833 800
Line 4 767 733 700
Line 5 667 633 600
Line 6 567 533 500
Line 7 467 433 400
Line 8 367 333 300
Line 9 267 233 200
Line 10 167 133 100
Period 1
Get your offer


Period 1 of 10: Decision
Your offer for Period 1 is:
933
Do you wish to accept the offer?
Remember: Whether you get the offer also depends on the period the computer selects.
Yes
No
Preview of the next period: Period 2
The next offer in Period 2 will be drawn from the numbers in the blue boxes below.
Line 1 1067 1033 1000
Line 2 967 933 900
Line 3 867 833 800
Line 4 767 733 700
Line 5 667 633 600
Line 6 567 533 500
Line 7 467 433 400
Line 8 367 333 300
Line 9 267 233 200
Line 10 167 133 100
Period 2

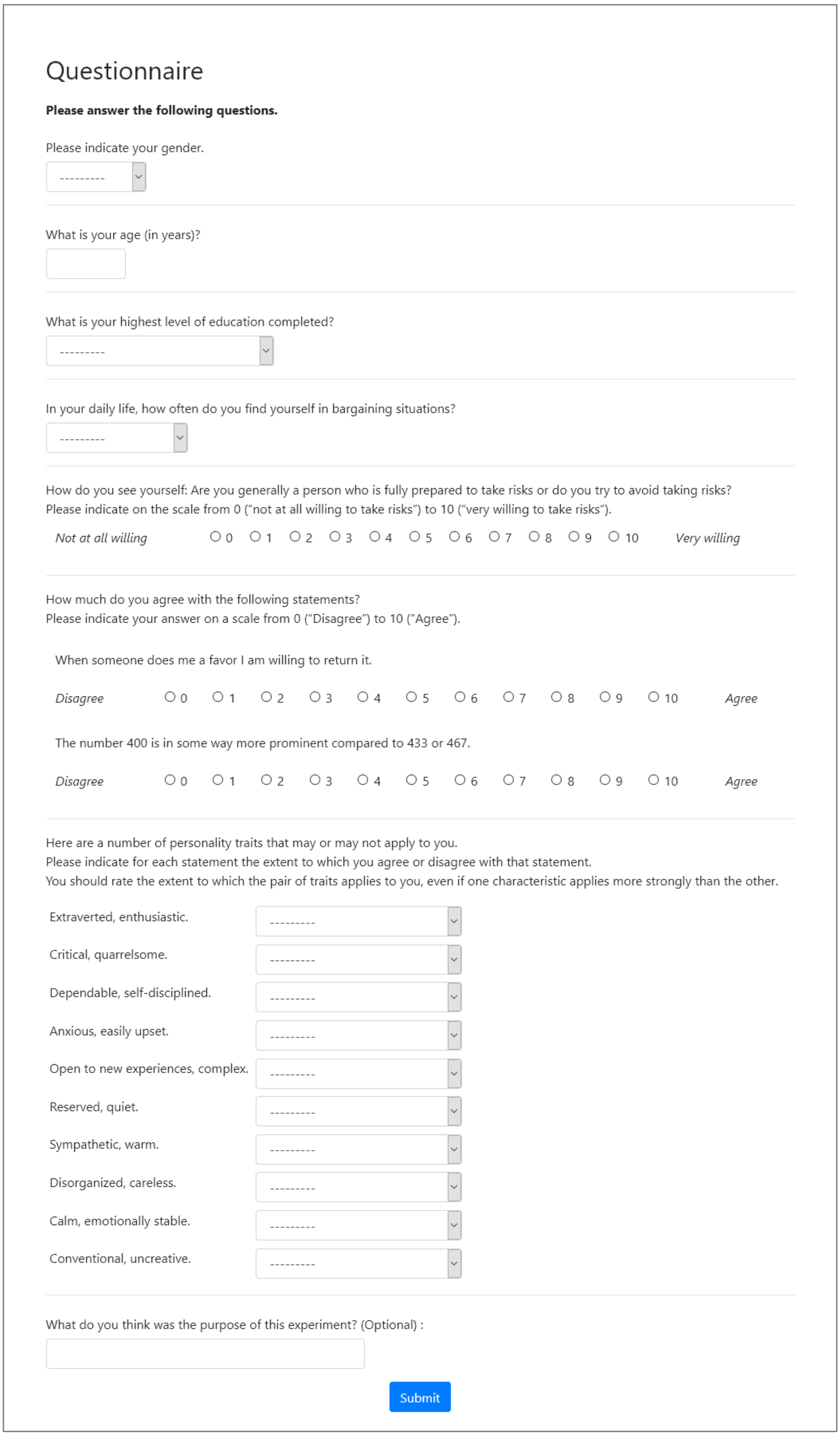

Period 2 of 10: Preview
Your offer for Period 2 will be drawn from the box below. But only the numbers in blue are available.
Line 1 1067 1033 1000
Line 2 967 933 900
Line 3 867 833 800
Line 4 767 733 700
Line 5 667 633 600
Line 6 567 533 500
Line 7 467 433 400
Line 8 367 333 300
Line 9 267 233 200
Line 10 167 133 100
Period 2
Get your offer


Period 2 of 10: Decision
Your offer for Period 2 is:
133
Do you wish to accept the offer?
Remember: Whether you get the offer also depends on the period the computer selects and the decision of your partner.
Yes
No
Preview of the next period: Period 3
The next offer in Period 3 will be drawn from the numbers in the blue boxes below.
Line 1 1067 1033 1000
Line 2 967 933 900
Line 3 867 833 800
Line 4 767 733 700
Line 5 667 633 600
Line 6 567 533 500
Line 7 467 433 400
Line 8 367 333 300
Line 9 267 233 200
Line 10 167 133 100
Period 3

Placeholder

Round 3 to Round 10

The above scheme is repeated.



Questionnaire

Please answer the following questions.

Please indicate your gender.

What is your age (in years)?

What is your highest level of education completed?

In your daily life, how often do you find yourself in bargaining situations?

How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?
Please indicate on the scale from 0 ("not at all willing to take risks") to 10 ("very willing to take risks").

Not at all willing ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 *Very willing*

How much do you agree with the following statements?
Please indicate your answer on a scale from 0 ("Disagree") to 10 ("Agree").

When someone does me a favor I am willing to return it.

Disagree ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 *Agree*

The number 400 is in some way more prominent compared to 433 or 467.

Disagree ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 *Agree*

Here are a number of personality traits that may or may not apply to you.
Please indicate for each statement the extent to which you agree or disagree with that statement.
You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

Extraverted, enthusiastic.	---------
Critical, quarrelsome.	---------
Dependable, self-disciplined.	---------
Anxious, easily upset.	---------
Open to new experiences, complex.	---------
Reserved, quiet.	---------
Sympathetic, warm.	---------
Disorganized, careless.	---------
Calm, emotionally stable.	---------
Conventional, uncreative.	---------

What do you think was the purpose of this experiment? (Optional) :

Submit

Thank you for participating!

This concludes the experiment. Please scroll down for the completion code.

Period	Offer	Your decision: Accept the offer?
1	933	Yes
2	133	No
3	300	No
4	167	Yes
5	667	Yes
6	567	Yes
7	300	Yes
8	167	Yes
9	200	Yes
10	167	Yes

The computer selected:

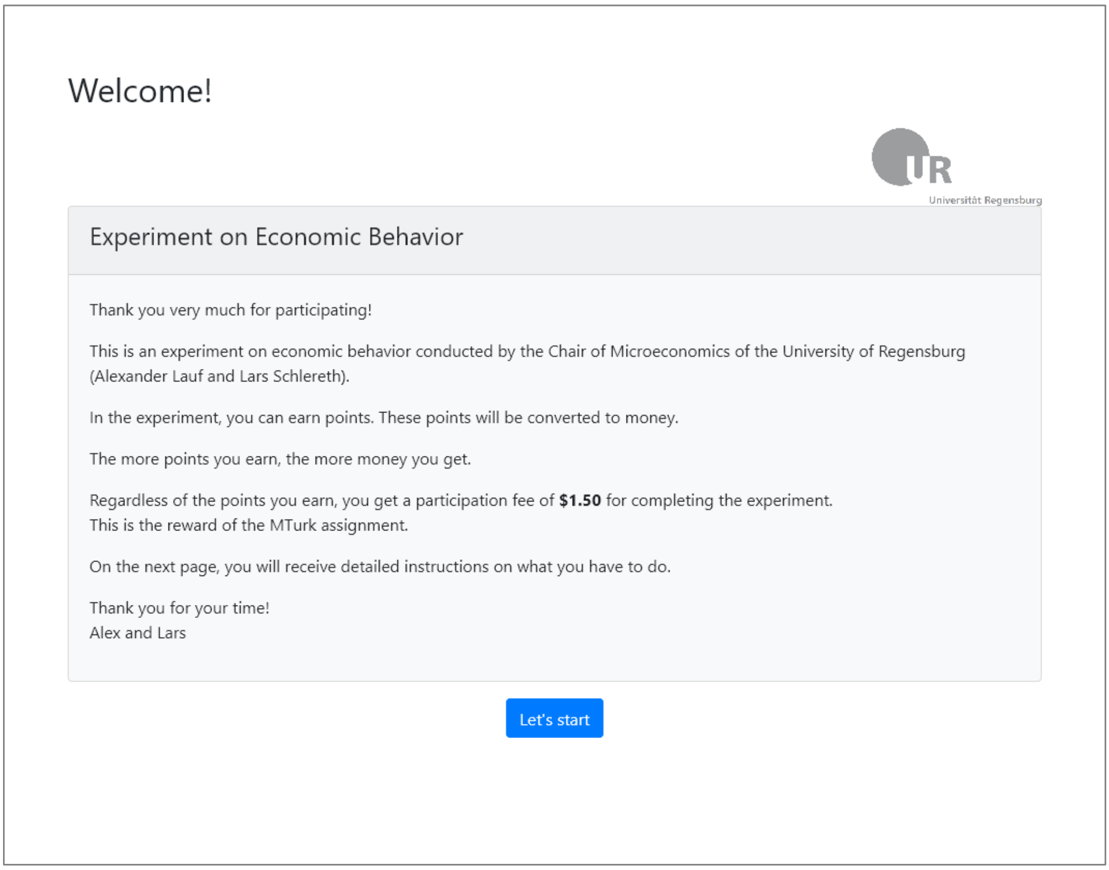
Period 8

The computer will determine your points by going from this period to Period 10 and check, which offer you accepted. Then the points will be converted to dollars and sent to you as bonus within the next days.

Your completion code:

03BF65

B.3.3 Experimental design: Partner



Welcome!

Universität Regensburg

Experiment on Economic Behavior

Thank you very much for participating!

This is an experiment on economic behavior conducted by the Chair of Microeconomics of the University of Regensburg (Alexander Lauf and Lars Schlereth).

In the experiment, you can earn points. These points will be converted to money.

The more points you earn, the more money you get.

Regardless of the points you earn, you get a participation fee of **$1.50** for completing the experiment.
This is the reward of the MTurk assignment.

On the next page, you will receive detailed instructions on what you have to do.

Thank you for your time!
Alex and Lars

Let's start

Instructions

What do you have to do?

In this experiment, you will make one decision in each of **10 periods**, i.e., a total of 10 decisions. In this experiment, you can earn points. After this experiment, the points will be converted into real money **(1000 Points = 1 Dollar)** and paid to you as a bonus later.

In the experiment, you will be matched with one **randomly selected player**, i.e., your 'partner'. The points you earn in the experiment depend on your own decisions and the decisions of your partner in a way explained below. You and your partner receive the **same instructions**.

In each period, you will receive an **offer**. Your partner will receive the same offer. The computer **randomly selects** one of the numbers in a blue box below. Each number is **equally likely** to be drawn. You can see the box for Period 1 below.

You can **accept** or **reject** the offer.

Line 1	1067	1033	1000
Line 2	967	933	900
Line 3	867	833	800
Line 4	767	733	700
Line 5	667	633	600
Line 6	567	533	500
Line 7	467	433	400
Line 8	367	333	300
Line 9	267	233	200
Line 10	167	133	100

Period 1

In each next period, **numbers are removed** from period to period. In Period 2, 'Line 1' is removed. In Period 3, additionally, 'Line 2' is removed. This is repeated until period 10, in which only the last line remains.

In each period, you decide to either **accept** or **reject** the offer. In subsequent periods, you can get a higher or lower offer, but keep in mind that the **three highest numbers** are removed in each period.

You can click on the right-arrow in the slider gallery below to see how these lines are removed.

Line 1	~~1067~~	~~1033~~	~~1000~~
Line 2	967	933	900
Line 3	867	833	800
Line 4	767	733	700
Line 5	667	633	600
Line 6	567	533	500
Line 7	467	433	400
Line 8	367	333	300
Line 9	267	233	200
Line 10	167	133	100

Period 2

In each period, you **privately** decide to either **accept or reject** the offer. Your partner also privately makes such decisions in each period. When you make your decisions, you do not know your partner's decisions and your partner does not know your decisions.

After you and your partner have made decisions for all 10 periods, your bonus in points is determined as follows:

The computer **randomly selects** a period out of Period 1 to 10. Then, the computer moves step-by-step from **this randomly selected period** to Period 10. The computer stops in the period, where for the first time **both of you have accepted** the respective period's offer. This offer is your bonus in points. Your partner gets the same bonus.

Therefore, you should play every period as if the computer selected the current period.

Your total payment in dollars is the sum of participation fee and your bonus in points converted to dollars (1000 Points = 1 Dollar):

$$\text{Payment} = \text{Participation Fee} + \frac{\text{Your Points}}{1000}$$

Now we ask you a few check-up questions. They do not have any impact on your payment. After you have finished the experiment, we will conduct a short survey.

Go to Check-Up Questions

Check-Up Question | Instructions

Check-Up Questions

Question (1)

The following questions have no effect on your amount of points. Please read carefully.

We want to check whether you understood the instructions. If you want to read the instructions again, you can find a tab 'Instructions' in the navigation bar above.

Suppose you are in **Period 1**. You got your offer from the numbers in the blue boxes below. You are about to move to the next period.

Line 1	1067	1033	1000
Line 2	967	933	900
Line 3	867	833	800
Line 4	767	733	700
Line 5	667	633	600
Line 6	567	533	500
Line 7	467	433	400
Line 8	367	333	300
Line 9	267	233	200
Line 10	167	133	100

Period 1

Which line is eliminated when you move to the next round (Period 2)?

- ○ Line 1: 1067, 1033, 1000
- ○ Line 5: 667, 633, 600
- ○ Line 9: 267, 233, 200
- ○ A random Line.

Instructions | Next Question

Check-Up Question | Instructions

Check-Up Question

Question (2)

The following questions have no effect on your amount of points. Please read carefully.

We want to check whether you understood the instructions. If you want to read the instructions again, you can find a tab 'Instructions' in the navigation bar above.

Suppose you have already made **all decisions**. Your partner has made all his or her decisions as well. In the picture below, you can see all of your decisions.

The computer has selected **Period 5**. As mentioned in the instructions, the computer starts in Period 5 and goes step-by-step to Period 10.

From which period do you get the offer as final payment?

- ○ Period 5
- ○ Period 10
- ○ Period 8
- ○ A random Period.

Instructions | Next Question

Check-Up Question | Instructions

Check-Up Question

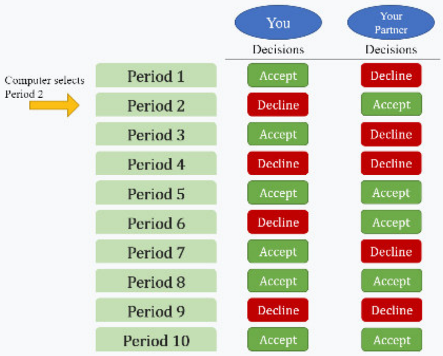
Question (3)

The following questions have no effect on your amount of points. Please read carefully.

We want to check whether you understood the instructions. If you want to read the instructions again, you can find a tab 'Instructions' in the navigation bar above.

Suppose you have already made **all decisions**. Your partner has made all his or her decisions as well. In the picture below, you can see all of your decisions.

The computer has selected **Period 2**. As mentioned in the instructions, the computer starts in Period 2 and goes step-by-step to Period 10.



From which period do you get the offer as final payment?

- ○ Period 5
- ○ Period 8
- ○ Period 10
- ○ A random Period.

Instructions | Next Question

Check-Up Question | Instructions

Check-Up Question

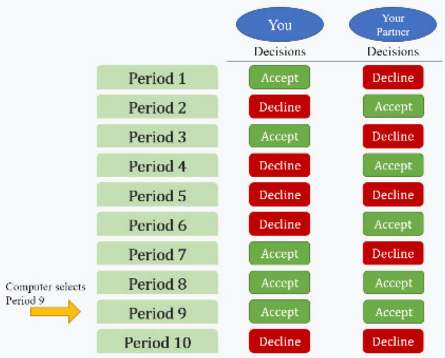
Question (4)

The following questions have no effect on your amount of points. Please read carefully.

We want to check whether you understood the instructions. If you want to read the instructions again, you can find a tab 'Instructions' in the navigation bar above.

Suppose you have already made **all decisions**. Your partner has made all his or her decisions as well. In the picture, below you can see all of your decisions.

The computer has selected **Period 9**. As mentioned in the instructions, the computer starts in Period 9 and goes step-by-step to Period 10.



From which period do you get the offer as final payment?

- ○ Period 2
- ○ Period 9
- ○ Period 6
- ○ A random Period.

Instructions | Continue

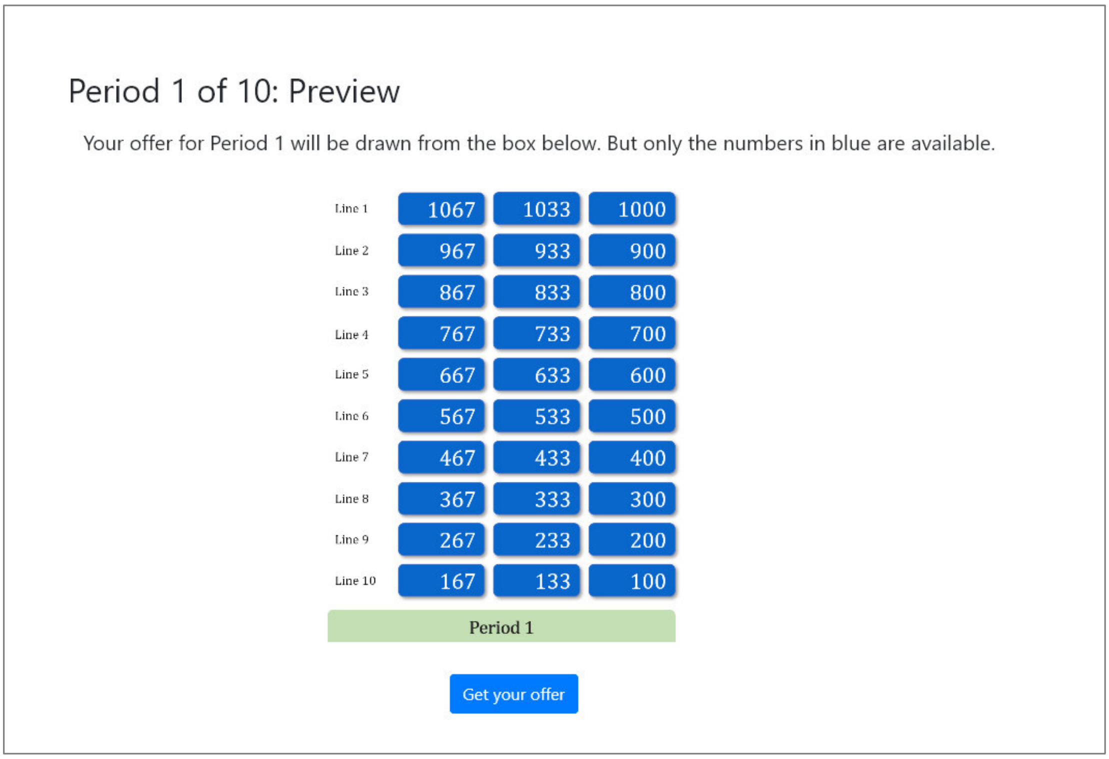
Start of the Experiment

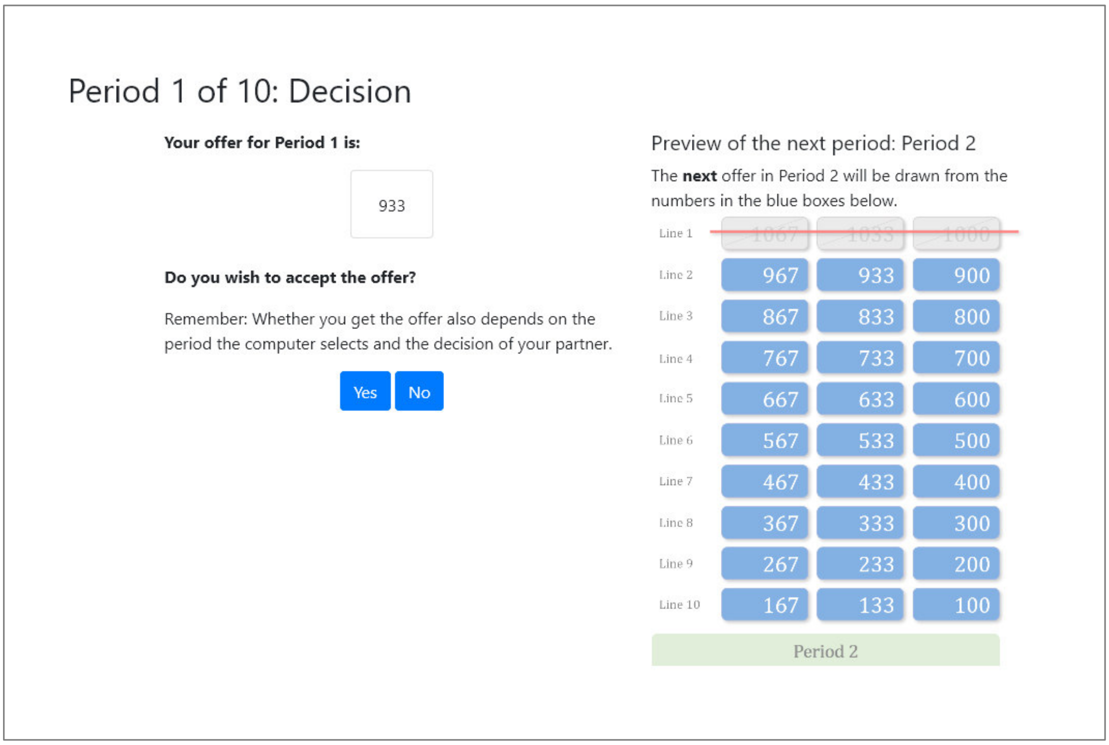
You have finished the check-up questions.

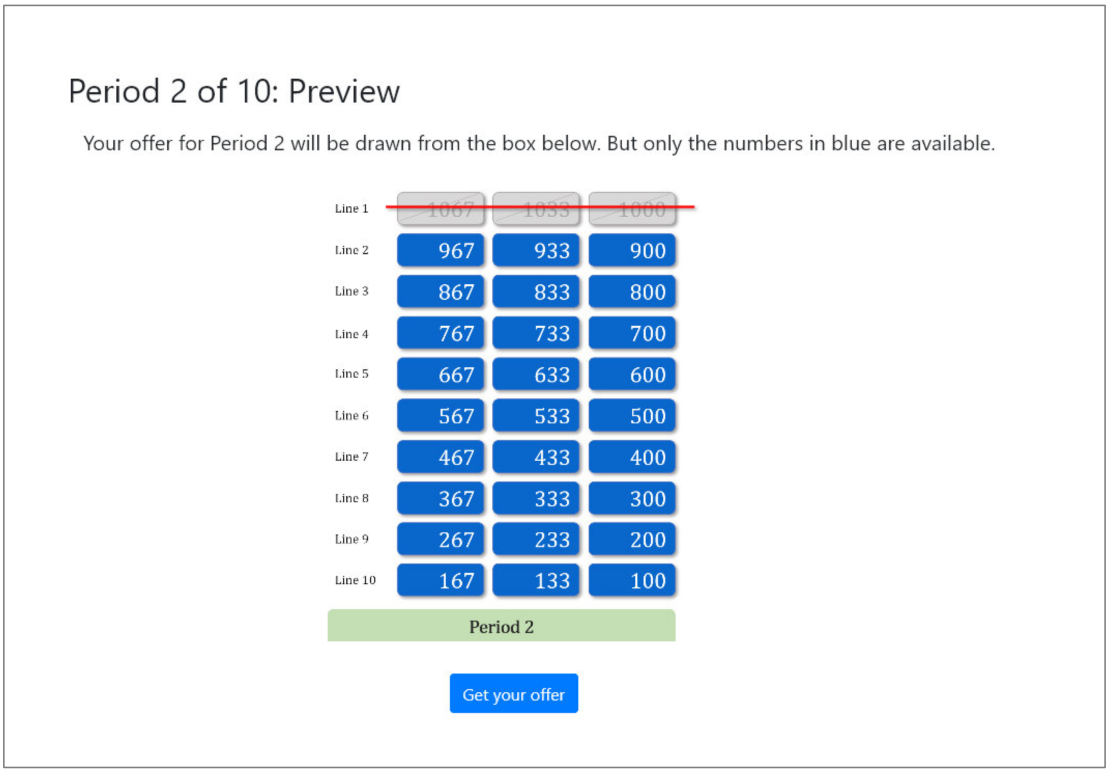
You can now start the experiment by clicking on the button below.

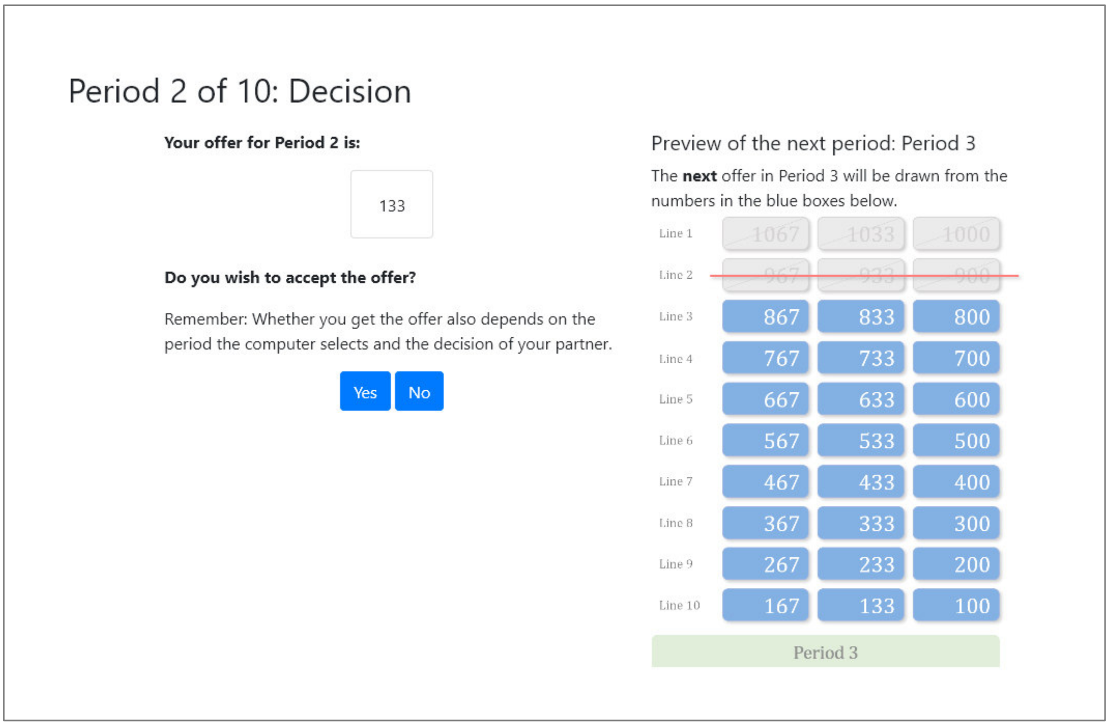
All your answers to the following questions will be relevant for your final payoff.

Start the Experiment!


Period 1 of 10: Preview
Your offer for Period 1 will be drawn from the box below. But only the numbers in blue are available.
Line 1 1067 1033 1000
Line 2 967 933 900
Line 3 867 833 800
Line 4 767 733 700
Line 5 667 633 600
Line 6 567 533 500
Line 7 467 433 400
Line 8 367 333 300
Line 9 267 233 200
Line 10 167 133 100
Period 1
Get your offer


Period 1 of 10: Decision
Your offer for Period 1 is:
933
Do you wish to accept the offer?
Remember: Whether you get the offer also depends on the period the computer selects and the decision of your partner.
Yes
No
Preview of the next period: Period 2
The next offer in Period 2 will be drawn from the numbers in the blue boxes below.
Line 1 1067 1033 1000
Line 2 967 933 900
Line 3 867 833 800
Line 4 767 733 700
Line 5 667 633 600
Line 6 567 533 500
Line 7 467 433 400
Line 8 367 333 300
Line 9 267 233 200
Line 10 167 133 100
Period 2


Period 2 of 10: Preview
Your offer for Period 2 will be drawn from the box below. But only the numbers in blue are available.
Line 1 1067 1033 1000
Line 2 967 933 900
Line 3 867 833 800
Line 4 767 733 700
Line 5 667 633 600
Line 6 567 533 500
Line 7 467 433 400
Line 8 367 333 300
Line 9 267 233 200
Line 10 167 133 100
Period 2
Get your offer


Period 2 of 10: Decision
Your offer for Period 2 is:
133
Do you wish to accept the offer?
Remember: Whether you get the offer also depends on the period the computer selects and the decision of your partner.
Yes
No
Preview of the next period: Period 3
The next offer in Period 3 will be drawn from the numbers in the blue boxes below.
Line 1 1067 1033 1000
Line 2 967 933 900
Line 3 867 833 800
Line 4 767 733 700
Line 5 667 633 600
Line 6 567 533 500
Line 7 467 433 400
Line 8 367 333 300
Line 9 267 233 200
Line 10 167 133 100
Period 3

Placeholder

Round 3 to Round 10

The above scheme is repeated.

Questionnaire

Please answer the following questions.

Please indicate your gender.

What is your age (in years)?

What is your highest level of education completed?

In your daily life, how often do you find yourself in bargaining situations?

How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?
Please indicate on the scale from 0 ("not at all willing to take risks") to 10 ("very willing to take risks").

Not at all willing ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 *Very willing*

How much do you agree with the following statements?
Please indicate your answer on a scale from 0 ("Disagree") to 10 ("Agree").

When someone does me a favor I am willing to return it.

Disagree ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 *Agree*

The number 400 is in some way more prominent compared to 433 or 467.

Disagree ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 *Agree*

Here are a number of personality traits that may or may not apply to you.
Please indicate for each statement the extent to which you agree or disagree with that statement.
You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

Extraverted, enthusiastic.	---------
Critical, quarrelsome.	---------
Dependable, self-disciplined.	---------
Anxious, easily upset.	---------
Open to new experiences, complex.	---------
Reserved, quiet.	---------
Sympathetic, warm.	---------
Disorganized, careless.	---------
Calm, emotionally stable.	---------
Conventional, uncreative.	---------

What do you think was the purpose of this experiment? (Optional) :

Submit

Thank you for participating!

This concludes the experiment. Please scroll down for the completion code.

Period	Offer	Your decision: Accept the offer?
1	933	Yes
2	133	Yes
3	300	Yes
4	167	Yes
5	667	Yes
6	567	Yes
7	300	Yes
8	167	Yes
9	200	Yes
10	167	Yes

The computer selected:

Period 8

The computer will determine your points by going from this period to Period 10 and check, which offer you and your partner jointly accepted. Then the points will be converted to dollars and sent to you as bonus within the next days.

Your completion code:

02FBC3

Appendix C

Appendix: Chapter 6

This appendix includes a detailed literature review, further tables, figures, and additional analyses for Chapter 6.

Ethics and Preregistration Statement This study received ethical approval from the UEBS Research Ethics Committee of the University of Exeter (Ethics application - eUEBS004241; 26.05.2021) and the Ethics Committee of the University of Regensburg (28.04.2021). It was preregistered on aspredicted.org (Nr. 68888) before data collection (see `Link` or Online (2022a)).

C.1 Summary statistics

Table C.1. Descriptives for the cisgender and transgender samples.

		Gender		
	Total	Cisgender	Transgender	p-value
	(N=780)	(N=425)	(N=355)	
Treatment				0.933
NEUTRAL	259 (33.2%)	143 (33.6%)	116 (32.7%)	
FEMININE	263 (33.7%)	141 (33.2%)	122 (34.4%)	
MASCULINE	258 (33.1%)	141 (33.2%)	117 (33.0%)	
Age (years)				0.516
Mean (SD)	24.4 (6.60)	24.3 (6.52)	24.6 (6.71)	
Height (cm)				0.002
Mean (SD)	170 (10.8)	171 (11.0)	169 (10.5)	
Student status				0.830
Yes	368 (47.2%)	202 (47.5%)	166 (46.8%)	
No	412 (52.8%)	223 (52.5%)	189 (53.2%)	
Highest education				0.094
University degree	266 (34.1%)	159 (37.4%)	107 (30.1%)	
High school diploma/A-levels	361 (46.3%)	189 (44.5%)	172 (48.5%)	
Other	153 (19.6%)	77 (18.1%)	76 (21.4%)	
Income: Less than 20,000 GBP				0.171
Yes	541 (69.4%)	286 (67.3%)	255 (71.8%)	
No	239 (30.6%)	139 (32.7%)	100 (28.2%)	
Religion				0.891
Non-religious	547 (70.1%)	295 (69.4%)	252 (71.0%)	
Religious	201 (25.8%)	112 (26.4%)	89 (25.1%)	
Not say	32 (4.1%)	18 (4.2%)	14 (3.9%)	
Residence				<0.001
Continental Europe	250 (32.1%)	169 (39.8%)	81 (22.8%)	
United Kingdom	205 (26.3%)	101 (23.8%)	104 (29.3%)	
United States	265 (34.0%)	133 (31.3%)	132 (37.2%)	
Other	60 (7.7%)	22 (5.2%)	38 (10.7%)	
BEM group:				0.002
Androgynous	188 (24.1%)	116 (27.3%)	72 (20.3%)	
Feminine	222 (28.5%)	104 (24.5%)	118 (33.2%)	
Masculine	151 (19.4%)	95 (22.4%)	56 (15.8%)	
Undifferentiated	219 (28.1%)	110 (25.9%)	109 (30.7%)	
BEM score: Feminine				0.730
Mean (SD)	41.8 (8.58)	41.8 (8.19)	41.7 (9.03)	
BEM score: Masculine				<0.001
Mean (SD)	33.9 (7.95)	35.0 (7.64)	32.5 (8.11)	
TCS				<0.001
Mean (SD)	3.67 (1.14)	4.47 (0.570)	2.71 (0.865)	
STT				<0.001
Mean (SD)	4.35 (4.59)	0.998 (1.47)	8.37 (3.76)	

Note: The table summarizes the characteristics of the cisgender and transgender samples. The education category other includes subjects that replied technical/community college, secondary education (e.g. GED/GCSE), no formal qualification, or don't know/not applicable. The religion category religious includes subjects that replied Buddhism, Christianity, Hinduism, Islam, Judaism, Paganism, Sikhism, or Spiritualism. The residence category other includes subjects that replied Australia, Canada, Chile, Israel, Japan, Mexico, New Zealand, or South Africa. The column p-value reports the p-values of χ^2-tests for categorical variables and the p-values of Wilcoxon-Mann Whitney tests for numerical variables between the cisgender and transgender column.

Table C.2. Descriptives by treatment for cismen.

		Treatment			
	Total	NEUTRAL	FEMININE	MASCULINE	*p*-value
Cismen					
	(N=214)	(N=72)	(N=71)	(N=71)	
Age (years)					0.042
Mean (SD)	24.1 (5.74)	25.8 (7.70)	24.1 (4.79)	22.5 (3.44)	
Height (cm)					0.449
Mean (SD)	178 (9.08)	180 (10.2)	177 (7.95)	177 (8.77)	
Student status					0.754
Yes	102 (47.7%)	32 (44.4%)	34 (47.9%)	36 (50.7%)	
No	112 (52.3%)	40 (55.6%)	37 (52.1%)	35 (49.3%)	
Highest education					0.237
University degree	72 (33.6%)	26 (36.1%)	27 (38.0%)	19 (26.8%)	
High school diploma/A-levels	94 (43.9%)	27 (37.5%)	28 (39.4%)	39 (54.9%)	
Other	48 (22.4%)	19 (26.4%)	16 (22.5%)	13 (18.3%)	
Income: Less than 20,000 GBP					0.841
Yes	135 (63.1%)	47 (65.3%)	43 (60.6%)	45 (63.4%)	
No	79 (36.9%)	25 (34.7%)	28 (39.4%)	26 (36.6%)	
Religion					0.820
Non-religious	144 (67.3%)	48 (66.7%)	47 (66.2%)	49 (69.0%)	
Religious	60 (28.0%)	21 (29.2%)	19 (26.8%)	20 (28.2%)	
Not say	10 (4.7%)	3 (4.2%)	5 (7.0%)	2 (2.8%)	
Residence					0.972
Continental Europe	95 (44.4%)	31 (43.1%)	30 (42.3%)	34 (47.9%)	
United Kingdom	50 (23.4%)	17 (23.6%)	17 (23.9%)	16 (22.5%)	
United States	65 (30.4%)	23 (31.9%)	23 (32.4%)	19 (26.8%)	
Other	4 (1.9%)	1 (1.4%)	1 (1.4%)	2 (2.8%)	
BEM group:					0.490
Androgynous	56 (26.2%)	19 (26.4%)	19 (26.8%)	18 (25.4%)	
Feminine	40 (18.7%)	15 (20.8%)	11 (15.5%)	14 (19.7%)	
Masculine	59 (27.6%)	22 (30.6%)	15 (21.1%)	22 (31.0%)	
Undifferentiated	59 (27.6%)	16 (22.2%)	26 (36.6%)	17 (23.9%)	
BEM score: Feminine					0.644
Mean (SD)	40.0 (8.55)	40.4 (8.52)	39.3 (9.13)	40.4 (8.03)	
BEM score: Masculine					0.522
Mean (SD)	35.7 (7.68)	36.2 (6.93)	34.9 (7.71)	36.0 (8.41)	
TCS					0.620
Mean (SD)	4.47 (0.591)	4.45 (0.541)	4.49 (0.590)	4.46 (0.645)	
STT					0.001
Mean (SD)	0.986 (1.46)	1.18 (1.09)	0.535 (0.939)	1.24 (2.01)	

Note: The table summarizes the characteristics of the cisgender and transgender samples. The education category other includes subjects that replied technical/community college, secondary education (e.g. GED/GCSE), no formal qualification, or don't know/not applicable. The religion category religious includes subjects that replied Buddhism, Christianity, Hinduism, Islam, Judaism, Paganism, Sikhism, or Spiritualism. The residence category other includes subjects that replied Australia, Canada, Chile, Israel, Japan, Mexico, New Zealand, or South Africa. The column *p*-value reports the *p*-values of χ^2-tests for categorical variables and the *p*-values of the Kruskal Wallis tests for numerical variables between the treatment columns.

Table C.3. Descriptives by treatment for ciswomen.

	Total	Treatment: NEUTRAL	Treatment: FEMININE	Treatment: MASCULINE	p-value
Ciswomen					
	(N=211)	(N=71)	(N=70)	(N=70)	
Age (years)					0.644
Mean (SD)	24.6 (7.23)	25.0 (7.83)	25.1 (7.79)	23.6 (5.90)	
Height (cm)					0.541
Mean (SD)	164 (7.96)	164 (9.88)	164 (7.02)	165 (6.60)	
Student status					0.813
Yes	100 (47.4%)	35 (49.3%)	34 (48.6%)	31 (44.3%)	
No	111 (52.6%)	36 (50.7%)	36 (51.4%)	39 (55.7%)	
Highest education					0.667
University degree	87 (41.2%)	32 (45.1%)	31 (44.3%)	24 (34.3%)	
High school diploma/A-levels	95 (45.0%)	29 (40.8%)	31 (44.3%)	35 (50.0%)	
Other	29 (13.7%)	10 (14.1%)	8 (11.4%)	11 (15.7%)	
Income: Less than 20,000 GBP					0.253
Yes	151 (71.6%)	53 (74.6%)	53 (75.7%)	45 (64.3%)	
No	60 (28.4%)	18 (25.4%)	17 (24.3%)	25 (35.7%)	
Religion					0.990
Non-religious	151 (71.6%)	51 (71.8%)	50 (71.4%)	50 (71.4%)	
Religious	52 (24.6%)	17 (23.9%)	17 (24.3%)	18 (25.7%)	
Not say	8 (3.8%)	3 (4.2%)	3 (4.3%)	2 (2.9%)	
Residence					0.589
Continental Europe	74 (35.1%)	28 (39.4%)	25 (35.7%)	21 (30.0%)	
United Kingdom	51 (24.2%)	19 (26.8%)	15 (21.4%)	17 (24.3%)	
United States	68 (32.2%)	21 (29.6%)	24 (34.3%)	23 (32.9%)	
Other	18 (8.5%)	3 (4.2%)	6 (8.6%)	9 (12.9%)	
BEM group:					0.187
Androgynous	60 (28.4%)	17 (23.9%)	28 (40.0%)	15 (21.4%)	
Feminine	64 (30.3%)	23 (32.4%)	19 (27.1%)	22 (31.4%)	
Masculine	36 (17.1%)	14 (19.7%)	11 (15.7%)	11 (15.7%)	
Undifferentiated	51 (24.2%)	17 (23.9%)	12 (17.1%)	22 (31.4%)	
BEM score: Feminine					0.212
Mean (SD)	43.5 (7.42)	43.3 (7.35)	44.7 (7.66)	42.6 (7.22)	
BEM score: Masculine					0.099
Mean (SD)	34.3 (7.54)	33.9 (7.31)	35.9 (8.18)	33.2 (6.93)	
TCS					0.878
Mean (SD)	4.48 (0.550)	4.55 (0.413)	4.47 (0.585)	4.42 (0.630)	
STT					0.906
Mean (SD)	1.01 (1.49)	1.15 (2.07)	0.957 (1.04)	0.914 (1.14)	

Note: The table summarizes the characteristics of the cisgender and transgender samples. The education category other includes subjects that replied technical/community college, secondary education (e.g. GED/GCSE), no formal qualification, or don't know/not applicable. The religion category religious includes subjects that replied Buddhism, Christianity, Hinduism, Islam, Judaism, Paganism, Sikhism, or Spiritualism. The residence category other includes subjects that replied Australia, Canada, Chile, Israel, Japan, Mexico, New Zealand, or South Africa. The column p-value reports the p-values of χ^2-tests for categorical variables and the p-values of the Kruskal Wallis tests for numerical variables between the treatment columns.

Table C.4. Descriptives by treatment for transmen.

	Total	Treatment: NEUTRAL	FEMININE	MASCULINE	*p*-value
Transmen					
	(N=215)	(N=72)	(N=72)	(N=71)	
Age (years)					0.775
Mean (SD)	24.3 (6.40)	25.1 (8.07)	24.0 (5.61)	23.7 (5.12)	
Height (cm)					0.301
Mean (SD)	164 (8.52)	165 (10.4)	164 (7.32)	164 (7.64)	
Student status					0.376
Yes	108 (50.2%)	34 (47.2%)	41 (56.9%)	33 (46.5%)	
No	107 (49.8%)	38 (52.8%)	31 (43.1%)	38 (53.5%)	
Highest education					0.891
University degree	63 (29.3%)	22 (30.6%)	19 (26.4%)	22 (31.0%)	
High school diploma/A-levels	109 (50.7%)	37 (51.4%)	39 (54.2%)	33 (46.5%)	
Other	43 (20.0%)	13 (18.1%)	14 (19.4%)	16 (22.5%)	
Income: Less than 20,000 GBP					0.355
Yes	155 (72.1%)	48 (66.7%)	52 (72.2%)	55 (77.5%)	
No	60 (27.9%)	24 (33.3%)	20 (27.8%)	16 (22.5%)	
Religion					0.892
Non-religious	144 (67.0%)	47 (65.3%)	49 (68.1%)	48 (67.6%)	
Religious	62 (28.8%)	23 (31.9%)	19 (26.4%)	20 (28.2%)	
Not say	9 (4.2%)	2 (2.8%)	4 (5.6%)	3 (4.2%)	
Residence					0.939
Continental Europe	47 (21.9%)	14 (19.4%)	17 (23.6%)	16 (22.5%)	
United Kingdom	64 (29.8%)	22 (30.6%)	23 (31.9%)	19 (26.8%)	
United States	85 (39.5%)	31 (43.1%)	25 (34.7%)	29 (40.8%)	
Other	19 (8.8%)	5 (6.9%)	7 (9.7%)	7 (9.9%)	
BEM group:					0.927
Androgynous	44 (20.5%)	17 (23.6%)	13 (18.1%)	14 (19.7%)	
Feminine	69 (32.1%)	20 (27.8%)	24 (33.3%)	25 (35.2%)	
Masculine	42 (19.5%)	16 (22.2%)	13 (18.1%)	13 (18.3%)	
Undifferentiated	60 (27.9%)	19 (26.4%)	22 (30.6%)	19 (26.8%)	
BEM score: Feminine					0.809
Mean (SD)	41.4 (8.90)	41.1 (8.18)	41.5 (9.47)	41.6 (9.13)	
BEM score: Masculine					0.597
Mean (SD)	33.6 (7.43)	34.0 (7.42)	33.1 (6.98)	33.6 (7.95)	
TCS					0.692
Mean (SD)	2.82 (0.868)	2.88 (0.946)	2.75 (0.857)	2.84 (0.800)	
STT					0.910
Mean (SD)	9.26 (3.15)	9.29 (3.50)	9.21 (3.01)	9.27 (2.96)	

Note: The table summarizes the characteristics of the cisgender and transgender samples. The education category other includes subjects that replied technical/community college, secondary education (e.g. GED/GCSE), no formal qualification, or don't know/not applicable. The religion category religious includes subjects that replied Buddhism, Christianity, Hinduism, Islam, Judaism, Paganism, Sikhism, or Spiritualism. The residence category other includes subjects that replied Australia, Canada, Chile, Israel, Japan, Mexico, New Zealand, or South Africa. The column *p*-value reports the *p*-values of χ^2-tests for categorical variables and the *p*-values of the Kruskal Wallis tests for numerical variables between the treatment columns.

Table C.5. Descriptives by treatment for transwomen.

		Treatment			
	Total	NEUTRAL	FEMININE	MASCULINE	*p*-value
Transwomen					
	(N=140)	(N=44)	(N=50)	(N=46)	
Age (years)					0.345
Mean (SD)	25.1 (7.15)	25.3 (5.91)	25.6 (9.01)	24.2 (5.89)	
Height (cm)					0.864
Mean (SD)	175 (10.1)	176 (8.41)	174 (13.4)	175 (7.24)	
Student status					0.939
Yes	58 (41.4%)	18 (40.9%)	20 (40.0%)	20 (43.5%)	
No	82 (58.6%)	26 (59.1%)	30 (60.0%)	26 (56.5%)	
Highest education					0.090
University degree	44 (31.4%)	13 (29.5%)	20 (40.0%)	11 (23.9%)	
High school diploma/A-levels	63 (45.0%)	16 (36.4%)	20 (40.0%)	27 (58.7%)	
Other	33 (23.6%)	15 (34.1%)	10 (20.0%)	8 (17.4%)	
Income: Less than 20,000 GBP					0.070
Yes	100 (71.4%)	37 (84.1%)	34 (68.0%)	29 (63.0%)	
No	40 (28.6%)	7 (15.9%)	16 (32.0%)	17 (37.0%)	
Religion					0.664
Non-religious	108 (77.1%)	33 (75.0%)	39 (78.0%)	36 (78.3%)	
Religious	27 (19.3%)	10 (22.7%)	10 (20.0%)	7 (15.2%)	
Not say	5 (3.6%)	1 (2.3%)	1 (2.0%)	3 (6.5%)	
Residence					0.257
Continental Europe	34 (24.3%)	11 (25.0%)	9 (18.0%)	14 (30.4%)	
United Kingdom	40 (28.6%)	11 (25.0%)	17 (34.0%)	12 (26.1%)	
United States	47 (33.6%)	13 (29.5%)	21 (42.0%)	13 (28.3%)	
Other	19 (13.6%)	9 (20.5%)	3 (6.0%)	7 (15.2%)	
BEM group:					0.333
Androgynous	28 (20.0%)	5 (11.4%)	12 (24.0%)	11 (23.9%)	
Feminine	49 (35.0%)	19 (43.2%)	17 (34.0%)	13 (28.3%)	
Masculine	14 (10.0%)	7 (15.9%)	3 (6.0%)	4 (8.7%)	
Undifferentiated	49 (35.0%)	13 (29.5%)	18 (36.0%)	18 (39.1%)	
BEM score: Feminine					0.973
Mean (SD)	42.3 (9.22)	42.8 (7.72)	41.3 (10.4)	42.8 (9.28)	
BEM score: Masculine					0.996
Mean (SD)	30.9 (8.83)	31.0 (9.17)	30.8 (8.75)	30.9 (8.80)	
TCS					0.745
Mean (SD)	2.54 (0.835)	2.54 (0.755)	2.55 (0.872)	2.52 (0.883)	
STT					0.027
Mean (SD)	7.01 (4.19)	7.93 (4.05)	5.80 (4.35)	7.46 (3.91)	

Note: The table summarizes the characteristics of the cisgender and transgender samples. The education category other includes subjects that replied technical/community college, secondary education (e.g. GED/GCSE), no formal qualification, or don't know/not applicable. The religion category religious includes subjects that replied Buddhism, Christianity, Hinduism, Islam, Judaism, Paganism, Sikhism, or Spiritualism. The residence category other includes subjects that replied Australia, Canada, Chile, Israel, Japan, Mexico, New Zealand, or South Africa. The column *p*-value reports the *p*-values of χ^2-tests for categorical variables and the *p*-values of the Kruskal Wallis tests for numerical variables between the treatment columns.

C.2 Priming (Part 1)

Fig. C.1 presents the number of marked words split up by treatments and subject groups. We do not find any differences in marked words within one priming condition across subject groups (KW, NEUTRAL: $p = 0.349$, FEMININE: $p = 0.874$, MASCULINE: $p = 0.112$). For the different subject groups separately across priming conditions, only the number of words marked by transmen didn't differ across priming conditions (KW, cismen: $p < 0.001$, ciswomen: $p = 0.038$, transmen: $p = 0.123$, transwomen: $p = 0.014$). Concerning gender differences in NEUTRAL, we do not see significant variations (MWU, $p = 0.820$). The same is true for sex differences (MWU, $p = 0.091$). As we did not pre-register to control for the number of words marked in our regressions, we do not add this variable in the reported analysis. However, please note that all main results remain qualitatively the same when we account for the heterogeneity in the number of marked words. The additional analyses are available on request.

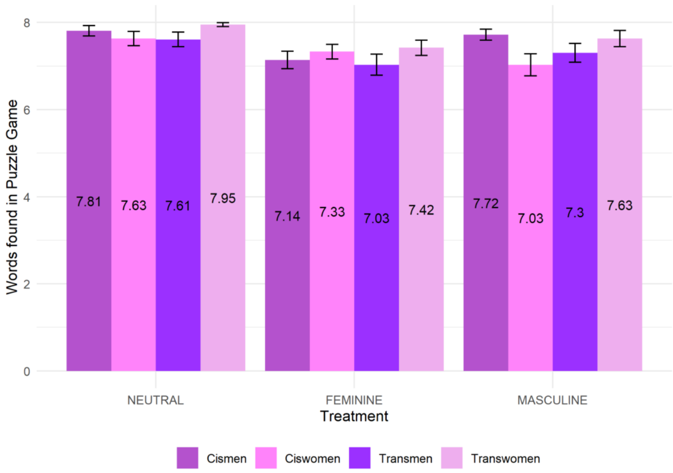


Figure C.1. Marked words in Part 1 by treatments and subject groups ($n = 780$).

Note: The bars show the average amount of marked words, and the error bars represent the standard error of the mean.

Table C.6. Words found in the priming task across treatments and subject groups.

Panel A: Priming across treatments

	Treatment			
Subject groups	NEUTRAL	FEMININE	MASCULINE	*p*-value
Cismen	7.806	7.141	7.718	<0.001
Ciswomen	7.634	7.329	7.029	0.038
Transmen	7.611	7.028	7.296	0.123
Transwomen	7.955	7.420	7.630	0.014

Panel B: Priming across subject groups

	Subject groups				
Treatment	Cismen	Ciswomen	Transmen	Transwomen	*p*-value
NEUTRAL	7.806	7.634	7.611	7.955	0.349
FEMININE	7.141	7.329	7.028	7.420	0.874
MASCULINE	7.718	7.029	7.296	7.630	0.112

Panel C: Priming across groups within NEUTRAL

	Group 1		Group 2		
	Subjects		Subjects		*p*-value
Case 1	Cisgender	7.720	Transgender	7.741	0.816
Case 2	Cismen	7.806	Ciswomen	7.634	0.339
Case 3	Transmen	7.611	Transwomen	7.955	0.122
Case 4	Female	7.622	Male	7.862	0.091
Case 5	Feminine	7.757	Masculine	7.708	0.820

Panel D: Priming in NEUTRAL compared to the other treatments

	NEUTRAL	FEMININE		MASCULINE	
Subject groups			*p*-value		*p*-value
Cismen	7.806	7.141	<0.001	7.718	0.345
Ciswomen	7.634	7.329	0.012	7.029	0.035
Transmen	7.611	7.028	0.040	7.296	0.207
Transwomen	7.955	7.420	0.004	7.630	0.100

Note: The column *p*-value reports the results of the tests performed column-wise. For continuous variables, we conducted Mann-Whitney U tests for two groups and Kruskal-Wallis tests for more than two groups.

C.3 Performance in the real effort math task (Part 2, 3, and 4)

The following tables summarize the performance in the math task by treatment and subject groups for Part 2 (Table C.7) and Part 3 (Table C.8). By treatments, ciswomen and cismen have differences in performance in MASCULINE in Part 2 (MWU; NEUTRAL: $p = 0.080$, FEMININE: $p = 0.205$, MASCULINE: $p = 0.037$) and across all treatments in Part 3 (MWU; NEUTRAL: $p = 0.004$, Part 3 FEMININE: $p = 0.010$, Part 3 MASCULINE: $p = 0.028$).

Transgender participants show performance differences in NEUTRAL in Part 2 and 3 (MWU, Part 2: NEUTRAL: $p = 0.007$, FEMININE: $p = 0.555$, MASCULINE: $p = 0.181$, Part 3: NEUTRAL: $p = 0.015$, FEMININE: $p = 0.600$, MASCULINE: $p = 0.053$). Concerning sex differences, male participants always have a higher performance than female ones in NEUTRAL and MASCULINE when facing piece-rate incentives (MWU, Part 2: NEUTRAL: $p = 0.003$, FEMININE: $p = 0.164$, MASCULINE: $p = 0.010$). Interestingly, this is true when they compete in Part 3 for all three treatments (MWU, NEUTRAL: $p < 0.001$, FEMININE: $p = 0.014$, MASCULINE: $p = 0.003$).

However, performances do not differ by the individual's gender (MWU, Part 2: NEUTRAL: $p = 0.755$, FEMININE: $p = 0.621$, MASCULINE: $p = 0.553$, Part 3: NEUTRAL: $p = 0.575$, FEMININE: $p = 0.161$, MASCULINE: $p = 0.675$). All differences vanish in Part 4, when we split up the data by those in the tournament (see the respective p-values in Table C.9 and Table C.9). For the priming intervention, we have no evidence of priming influencing the performance, independent of the part or subject group (KW, cismen: $p's > 0.478$, ciswomen: $p's > 0.562$, transmen: $p's > 0.956$, or transwomen: $p'S > 0.170$).

Please note that we can not exclude that the math task is not influenced by a participant's gender and sex, combinations of it, in addition to interactions with priming. However, we can control how performance heterogeneity affects competitiveness by adding individual performances to our regressions measuring competitiveness. See Table C.15 to Table C.16.

C.3.1 Performance Part 2

Table C.7. Performance in Part 2 across treatments and subject groups.

Panel A: Performance in Part 2 across treatments

	Treatment			
Subject groups	NEUTRAL	FEMININE	MASCULINE	p-value
Cismen	8.458	8.704	9.423	0.478
Ciswomen	7.535	7.929	8.014	0.719
Transmen	7.750	7.778	7.718	0.956
Transwomen	9.409	8.740	8.739	0.285

Panel B: Performance in Part 2 across subject groups

	Subject groups				
Treatment	Cismen	Ciswomen	Transmen	Transwomen	p-value
NEUTRAL	8.458	7.535	7.750	9.409	0.011
FEMININE	8.704	7.929	7.778	8.740	0.529
MASCULINE	9.423	8.014	7.718	8.739	0.062

Panel C: Performance in Part 2 across groups within NEUTRAL

	Group 1		Group 2		
	Subjects		Subjects		p-value
Case 1	Cisgender	8.000	Transgender	8.379	0.405
Case 2	Cismen	8.458	Ciswomen	7.535	0.080
Case 3	Transmen	7.750	Transwomen	9.409	0.007
Case 4	Female	7.643	Male	8.819	0.003
Case 5	Feminine	8.252	Masculine	8.104	0.755

Panel D: Performance in Part 2 in NEUTRAL compared to the other treatments

	NEUTRAL	FEMININE		MASCULINE	
Subject groups			p-value		p-value
Cismen	8.458	8.704	0.797	9.423	0.231
Ciswomen	7.535	7.929	0.488	8.014	0.476
Transmen	7.750	7.778	0.832	7.718	0.932
Transwomen	9.409	8.740	0.125	8.739	0.255

Note: The column p-value reports the results of the tests performed column-wise. For continuous variables, we conducted Mann-Whitney U tests for two groups and Kruskal-Wallis tests for more than two groups.

C.3.2 Performance Part 3

Table C.8. Performance in Part 3 across treatments and subject groups.

Panel A: Performance in Part 3 across treatments

	Treatment			
Subject groups	NEUTRAL	FEMININE	MASCULINE	p-value
Cismen	9.333	10.070	9.930	0.593
Ciswomen	7.423	7.957	8.271	0.562
Transmen	7.833	7.736	7.930	0.979
Transwomen	9.659	8.500	9.326	0.170

Panel B: Performance in Part 3 across subject groups

	Subject groups				
Treatment	Cismen	Ciswomen	Transmen	Transwomen	p-value
NEUTRAL	9.333	7.423	7.833	9.659	0.002
FEMININE	10.070	7.957	7.736	8.500	0.021
MASCULINE	9.930	8.271	7.930	9.326	0.024

Panel C: Performance in Part 3 across groups within NEUTRAL

	Group 1		Group 2		
	Subjects		Subjects		p-value
Case 1	Cisgender	8.385	Transgender	8.526	0.612
Case 2	Cismen	9.333	Ciswomen	7.423	0.004
Case 3	Transmen	7.833	Transwomen	9.659	0.015
Case 4	Female	7.629	Male	9.457	<0.001
Case 5	Feminine	8.278	Masculine	8.583	0.575

Panel D: Performance in Part 3 in NEUTRAL compared to the other treatments

	NEUTRAL	FEMININE		MASCULINE	
Subject groups			p-value		p-value
Cismen	9.333	10.070	0.353	9.930	0.406
Ciswomen	7.423	7.957	0.396	8.271	0.325
Transmen	7.833	7.736	0.920	7.930	0.832
Transwomen	9.659	8.500	0.081	9.326	0.686

Note: The column p-value reports the results of the tests performed column-wise. For continuous variables, we conducted Mann-Whitney U tests for two groups and Kruskal-Wallis tests for more than two groups.

C.3.3 Performance Part 4

Table C.9. Performance in Part 4 of *competing* subjects across treatments and subject groups.

Panel A: Performance in Part 4 of competing subjects across treatments

	Treatment			
Subject groups	NEUTRAL	FEMININE	MASCULINE	p-value
Cismen	9.500	10.321	9.312	0.765
Ciswomen	8.391	8.947	9.053	0.763
Transmen	9.000	8.211	8.474	0.839
Transwomen	9.429	8.333	9.167	0.670

Panel B: Performance in Part 4 of competing subjects across subject groups

	Subject groups				
Treatment	Cismen	Ciswomen	Transmen	Transwomen	p-value
NEUTRAL	9.500	8.391	9.000	9.429	0.705
FEMININE	10.321	8.947	8.211	8.333	0.389
MASCULINE	9.312	9.053	8.474	9.167	0.923

Panel C: Performance in Part 4 of competing subjects across groups within NEUTRAL

	Group 1		Group 2		
	Subjects		Subjects		p-value
Case 1	Cisgender	8.933	Transgender	9.176	0.651
Case 2	Cismen	9.500	Ciswomen	8.391	0.278
Case 3	Transmen	9.000	Transwomen	9.429	0.860
Case 4	Female	8.674	Male	9.472	0.358
Case 5	Feminine	8.784	Masculine	9.262	0.475

Panel D: Performance in Part 4 of competing subjects in NEUTRAL compared to the other treatments

	NEUTRAL	FEMININE		MASCULINE	
Subject groups			p-value		p-value
Cismen	9.500	10.321	0.556	9.312	0.965
Ciswomen	8.391	8.947	0.638	9.053	0.494
Transmen	9.000	8.211	0.563	8.474	0.671
Transwomen	9.429	8.333	0.360	9.167	0.797

Note: The column p-value reports the results of the tests performed column-wise. For continuous variables, we conducted Mann-Whitney U tests for two groups and Kruskal-Wallis tests for more than two groups.

Table C.10. Performance in Part 4 of *non-competing* subjects across treatments and subject groups.

Panel A: Performance in Part 4 of non-competing subjects across treatments

	Treatment			
Subject groups	NEUTRAL	FEMININE	MASCULINE	p-value
Cismen	8.960	9.093	9.026	0.994
Ciswomen	7.979	9.098	8.725	0.341
Transmen	8.712	8.453	8.288	0.825
Transwomen	8.800	8.906	9.500	0.703

Panel B: Performance in Part 4 of non-competing subjects across subject groups

	Subject groups				
Treatment	Cismen	Ciswomen	Transmen	Transwomen	p-value
NEUTRAL	8.960	7.979	8.712	8.800	0.577
FEMININE	9.093	9.098	8.453	8.906	0.929
MASCULINE	9.026	8.725	8.288	9.500	0.376

Panel C: Performance in Part 4 of non-competing subjects across groups within NEUTRAL

	Group 1		Group 2		
	Subjects		Subjects		p-value
Case 1	Cisgender	8.480	Transgender	8.744	0.874
Case 2	Cismen	8.960	Ciswomen	7.979	0.159
Case 3	Transmen	8.712	Transwomen	8.800	0.977
Case 4	Female	8.360	Male	8.900	0.292
Case 5	Feminine	8.295	Masculine	8.833	0.309

Panel D: Performance in Part 4 of non-competing subjects in NEUTRAL compared to the other treatments

	NEUTRAL	FEMININE		MASCULINE	
Subject groups			p-value		p-value
Cismen	8.960	9.093	0.914	9.026	0.960
Ciswomen	7.979	9.098	0.135	8.725	0.529
Transmen	8.712	8.453	0.921	8.288	0.606
Transwomen	8.800	8.906	0.843	9.500	0.445

Note: The column p-value reports the results of the tests performed column-wise. For continuous variables, we conducted Mann-Whitney U tests for two groups and Kruskal-Wallis tests for more than two groups.

C.4 Beliefs (Part 3)

C.4.1 Non-parametric tests

Table C.11. Beliefs in Part 3 across treatments and subject groups.

Panel A: Beliefs in Part 3 across treatments

	Treatment			
Subject groups	NEUTRAL	FEMININE	MASCULINE	p-value
Cismen	2.139	1.944	2.070	0.391
Ciswomen	2.606	2.500	2.571	0.793
Transmen	2.542	2.653	2.704	0.633
Transwomen	2.205	2.440	2.304	0.396

Panel B: Beliefs in Part 3 across subject groups

	Subject groups				
Treatment	Cismen	Ciswomen	Transmen	Transwomen	p-value
NEUTRAL	2.139	2.606	2.542	2.205	0.006
FEMININE	1.944	2.500	2.653	2.440	<0.001
MASCULINE	2.070	2.571	2.704	2.304	<0.001

Panel C: Beliefs in Part 3 across groups within NEUTRAL

	Group 1		Group 2		
	Subjects		Subjects		p-value
Case 1	Cisgender	2.371	Transgender	2.414	0.746
Case 2	Cismen	2.139	Ciswomen	2.606	0.003
Case 3	Transmen	2.542	Transwomen	2.205	0.061
Case 4	Female	2.573	Male	2.164	0.001
Case 5	Feminine	2.452	Masculine	2.340	0.362

Panel D: Beliefs in Part 3 in NEUTRAL compared to the other treatments

	NEUTRAL	FEMININE		MASCULINE	
Subject groups			p-value		p-value
Cismen	2.139	1.944	0.177	2.070	0.567
Ciswomen	2.606	2.500	0.496	2.571	0.772
Transmen	2.542	2.653	0.529	2.704	0.354
Transwomen	2.205	2.440	0.178	2.304	0.537

Note: The column p-value reports the results of the tests performed column-wise. For continuous variables, we conducted Mann-Whitney U tests for two groups and Kruskal-Wallis tests for more than two groups.

C.4.2 Regressions

Table C.12. OLS regression for NEUTRAL. Dependent variable: Beliefs in Part 3.

	(1)	(2)	(3)
Ciswomen	0.467 **	0.586 **	0.494 **
	(0.155)	(0.176)	(0.156)
Transmen	0.403 *	0.511 **	-0.155
	(0.159)	(0.180)	(0.322)
Transwomen	0.066	0.194	-0.551
	(0.172)	(0.173)	(0.324)
Age		0.000	
		(0.009)	
Height		0.005	
		(0.006)	
Student status		-0.169	
		(0.136)	
Income: < 20,000 GBP		-0.077	
		(0.135)	
Religion: Religious		0.243	
		(0.135)	
Religion: Not say		0.460	
		(0.313)	
Residence: US		-0.080	
		(0.152)	
Residence: UK		0.097	
		(0.170)	
Residence: Other		-0.408 *	
		(0.202)	
TCS			-0.253 **
			(0.093)
STT			0.020
			(0.022)
Const.	2.139 ***	1.226	3.242 ***
	(0.105)	(1.195)	(0.436)
N	259	259	259
Adj. R2	0.035	0.057	0.057
H_0: Sex	0.001	0.001	0.000
H_0: Gender	0.587	0.259	0.682

Note: The beliefs in Part 3 are the participants' belief about how their performance ranks within the group (1 = best to 4 = worst). Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious cisman, who earns more than 20K GBP, and lives in continental Europe. In the last column from the right, the baseline is a cisman. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. H_0: Sex reports the p-values of a joint coefficient test comparing Male (Cismen and Transwomen) with Female (Ciswomen and Transmen). H_0: Gender reports the p-values of a joint coefficient test comparing Masculine (Cismen and Transmen) with Feminine (Ciswomen and Transwomen).

Table C.13. OLS regression for all treatments. Dependent variable: Beliefs in Part 3.

	(1)	(2)	(3)
Ciswomen	0.467 **	0.444 **	0.484 **
	(0.155)	(0.162)	(0.155)
Transmen	0.403 *	0.405 *	0.093
	(0.159)	(0.164)	(0.229)
Transwomen	0.066	0.080	-0.291
	(0.172)	(0.174)	(0.234)
Treatment: FEMININE	-0.195	-0.187	-0.184
	(0.149)	(0.146)	(0.148)
Treatment: MASCULINE	-0.068	-0.045	-0.067
	(0.151)	(0.148)	(0.150)
FEMININE x Ciswomen	0.090	0.094	0.065
	(0.218)	(0.216)	(0.217)
MASCULINE x Ciswomen	0.034	0.034	0.013
	(0.220)	(0.215)	(0.219)
FEMININE x Transmen	0.306	0.309	0.273
	(0.218)	(0.216)	(0.215)
MASCULINE x Transmen	0.231	0.210	0.222
	(0.222)	(0.220)	(0.219)
FEMININE x Transwomen	0.431	0.421	0.434
	(0.239)	(0.238)	(0.236)
MASCULINE x Transwomen	0.168	0.145	0.166
	(0.242)	(0.242)	(0.241)
Const.	2.139 ***	2.547 ***	2.860 ***
	(0.105)	(0.693)	(0.250)
Controls (Age, Height, Student status, Income, Religion, Residence)	–	Yes	–
Controls (TCS, STT)	–	–	Yes
N	780	780	780
Adj. R2	0.055	0.074	0.067
H_0: FEMININE on Cismen	0.192	0.201	0.213
H_0: MASCULINE on Cismen	0.651	0.762	0.654
H_0: FEMININE on Ciswomen	0.504	0.553	0.454
H_0: MASCULINE on Ciswomen	0.830	0.943	0.736
H_0: FEMININE on Transmen	0.483	0.443	0.568
H_0: MASCULINE on Transmen	0.317	0.309	0.331
H_0: FEMININE on Transwomen	0.207	0.211	0.178
H_0: MASCULINE on Transwomen	0.597	0.601	0.599

Note: The beliefs in Part 3 are the participants' belief about how their performance ranks within the group (1 = best to 4 = worst). Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious cisman, who earns more than 20K GBP, and lives in continental Europe. In the last column from the right, the baseline is a cisman. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Rows starting with H_0 report the p-values of a joint coefficient test that the coefficients' sum equals 0. For example, "H_0: FEMININE on Ciswomen" tests the effect of the treatment (FEMININE) on the subject group (Ciswomen).

C.5 Competitiveness (Part 4)

C.5.1 Bar graphs

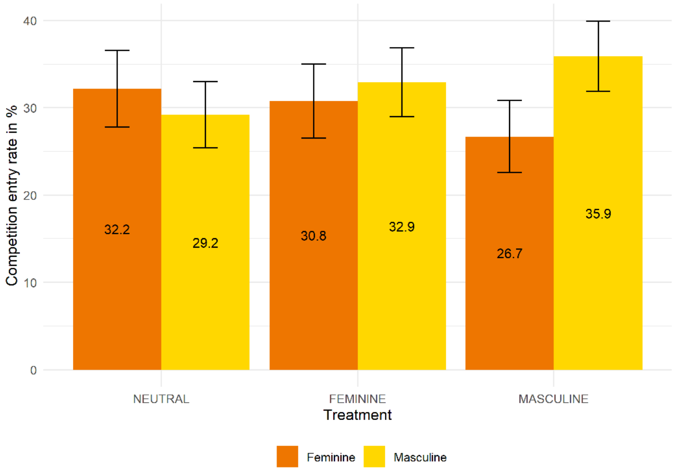


Figure C.2. Tournament entry rates in Part 4 by treatments and gender ($n = 780$).

Note: The bars show the percentage of participants (between 0 and 100) who chose to compete rather than to perform under piece–rate incentives. The error bars represent the standard error of the mean.

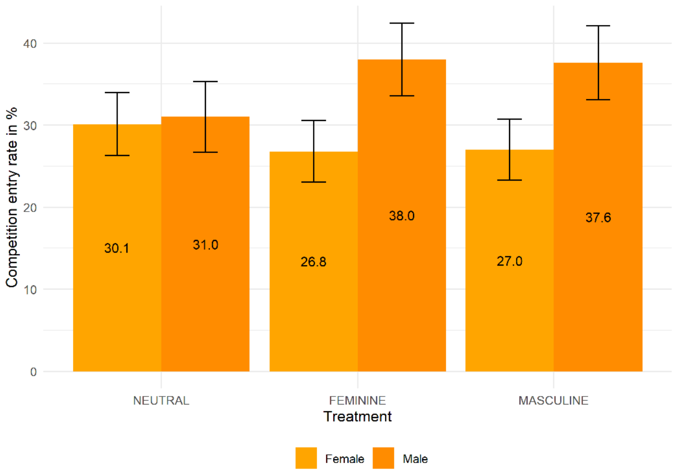


Figure C.3. Tournament entry rates in Part 4 by treatments and sex ($n = 780$).

Note: The bars show the percentage of participants (between 0 and 100) who chose to compete rather than to perform under piece–rate incentives. The error bars represent the standard error of the mean.

C.5.2 Non-parametric tests

Table C.14. Tournament entry rates across treatments and subject groups.

Panel A: Competitiveness across treatments

	Treatment			
Subject groups	NEUTRAL	FEMININE	MASCULINE	p-value
Cismen	30.6%	39.4%	45.1%	0.198
Ciswomen	32.4%	27.1%	27.1%	0.729
Transmen	27.8%	26.4%	26.8%	0.981
Transwomen	31.8%	36.0%	26.1%	0.578

Panel B: Competitiveness across subject groups

	Subject groups				
Treatment	Cismen	Ciswomen	Transmen	Transwomen	p-value
NEUTRAL	30.6%	32.4%	27.8%	31.8%	0.939
FEMININE	39.4%	27.1%	26.4%	36.0%	0.264
MASCULINE	45.1%	27.1%	26.8%	26.1%	0.046

Panel C: Competitiveness across groups within NEUTRAL

	Group 1		Group 2		
	Subjects		Subjects		p-value
Case 1	Cisgender	31.5%	Transgender	29.3%	0.708
Case 2	Cismen	30.6%	Ciswomen	32.4%	0.813
Case 3	Transmen	27.8%	Transwomen	31.8%	0.643
Case 4	Female	30.1%	Male	31.0%	0.867
Case 5	Feminine	32.2%	Masculine	29.2%	0.601

Panel D: Competitiveness in NEUTRAL compared to the other treatments

	NEUTRAL	FEMININE		MASCULINE	
Subject groups			p-value		p-value
Cismen	30.6%	39.4%	0.265	45.1%	0.073
Ciswomen	32.4%	27.1%	0.495	27.1%	0.495
Transmen	27.8%	26.4%	0.851	26.8%	0.891
Transwomen	31.8%	36.0%	0.669	26.1%	0.549

Note: The column p-value reports the results of χ^2 tests performed column-wise.

C.5.3 Regressions

Table C.15. Probit regression for NEUTRAL. Dependent variable: Competition.

	(1)	(2)	(3)	(4)	(5)
Ciswomen	0.052	0.173	0.281	0.360	0.268
	(0.220)	(0.231)	(0.254)	(0.297)	(0.254)
Transmen	-0.081	0.013	0.106	0.114	0.516
	(0.223)	(0.231)	(0.239)	(0.292)	(0.462)
Transwomen	0.036	0.108	0.084	-0.076	0.478
	(0.252)	(0.261)	(0.265)	(0.306)	(0.472)
Perf. tournament		0.003	-0.064 *	-0.054	-0.064 *
		(0.025)	(0.029)	(0.030)	(0.029)
Delta perf.		0.122 ***	0.096 *	0.103 **	0.095 *
		(0.035)	(0.039)	(0.038)	(0.040)
Belief tournament			-0.522 ***	-0.516 ***	-0.518 ***
			(0.123)	(0.127)	(0.124)
Risk			0.124	0.137	0.130
			(0.074)	(0.077)	(0.074)
Age				0.026 *	
				(0.012)	
Height				0.007	
				(0.010)	
Student status				-0.073	
				(0.209)	
Income: < 20,000 GBP				0.442	
				(0.232)	
Religion: Religious				0.081	
				(0.207)	
Religion: Not say				0.268	
				(0.485)	
Residence: US				0.159	
				(0.254)	
Residence: UK				0.421	
				(0.270)	
Residence: Other				0.564	
				(0.391)	
TCS					0.094
					(0.139)
STT					-0.033
					(0.033)
Const.	-0.508 **	-0.670 *	0.823	-1.618	0.419
	(0.156)	(0.272)	(0.468)	(2.000)	(0.789)
N	259	259	259	259	259
Pseudo R2 (McFadden)	0.001	0.059	0.129	0.168	0.132
H_0: Sex	0.846	0.823	0.415	0.214	0.420
H_0: Gender	0.614	0.437	0.470	0.656	0.527

Note: Competition is a binary variable equal to 1 if the participant enters the tournament in Part 4 and 0 otherwise. Delta perf. is the difference in performance between Part 3 (tournament) and Part 2 (piece-rate). Belief tournament is the participants' belief of their performance rank within their group in Part 3, where the value 1 represents the rank with the highest performance. Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious cisman, who earns more than 20K GBP, and lives in continental Europe. In the last column from the right, the baseline is a cisman. *** $p <$ 0.001; ** $p < 0.01$; * $p < 0.05$. H_0: Sex reports the p-values of a joint coefficient test comparing Male (Cismen and Transwomen) with Female (Ciswomen and Transmen). H_0: Gender reports the p-values of a joint coefficient test comparing Masculine (Cismen and Transmen) with Feminine (Ciswomen and Transwomen).

Table C.16. Probit regression for all treatments. Dependent variable: Competition.

	(1)	(2)	(3)	(4)	(5)
Ciswomen	0.052	0.167	0.307	0.461	0.280
	(0.220)	(0.229)	(0.255)	(0.270)	(0.255)
Transmen	-0.081	0.008	0.127	0.189	0.634
	(0.223)	(0.230)	(0.241)	(0.270)	(0.343)
Transwomen	0.036	0.122	0.078	0.079	0.635
	(0.252)	(0.262)	(0.272)	(0.287)	(0.359)
Treatment: FEMININE	0.241	0.187	0.129	0.167	0.104
	(0.218)	(0.223)	(0.227)	(0.241)	(0.231)
Treatment: MASCULINE	0.385	0.473 *	0.428	0.544 *	0.423
	(0.217)	(0.225)	(0.232)	(0.246)	(0.234)
FEMININE x Ciswomen	-0.392	-0.339	-0.318	-0.383	-0.285
	(0.313)	(0.329)	(0.360)	(0.365)	(0.360)
MASCULINE x Ciswomen	-0.536	-0.678 *	-0.645	-0.755 *	-0.614
	(0.312)	(0.323)	(0.342)	(0.348)	(0.343)
FEMININE x Transmen	-0.282	-0.201	-0.107	-0.112	-0.052
	(0.313)	(0.325)	(0.329)	(0.340)	(0.333)
MASCULINE x Transmen	-0.415	-0.543	-0.405	-0.475	-0.369
	(0.313)	(0.319)	(0.327)	(0.341)	(0.333)
FEMININE x Transwomen	-0.126	-0.018	0.187	0.131	0.150
	(0.347)	(0.352)	(0.357)	(0.366)	(0.360)
MASCULINE x Transwomen	-0.552	-0.690	-0.573	-0.651	-0.583
	(0.356)	(0.366)	(0.377)	(0.393)	(0.379)
Perf. tournament		-0.008	-0.067 ***	-0.066 ***	-0.066 ***
		(0.014)	(0.016)	(0.016)	(0.016)
Delta perf.		0.136 ***	0.108 ***	0.114 ***	0.110 ***
		(0.020)	(0.021)	(0.022)	(0.021)
Belief tournament			-0.561 ***	-0.557 ***	-0.543 ***
			(0.071)	(0.072)	(0.071)
Risk			0.150 ***	0.153 ***	0.153 ***
			(0.045)	(0.046)	(0.045)
Const.	-0.508 **	-0.587 **	0.865 **	-1.757	-0.072
	(0.156)	(0.200)	(0.310)	(1.255)	(0.480)
Controls (Age, Height, Student status, Income, Religion, Residence)	–	–	–	Yes	–
Controls (TCS, STT)	–	–	–	–	Yes
N	780	780	780	780	780
Pseudo R2 (McFadden)	0.013	0.082	0.161	0.178	0.167
H_0: FEMININE on Cismen	0.266	0.402	0.579	0.478	0.656
H_0: MASCULINE on Cismen	0.074	0.034	0.063	0.021	0.068
H_0: FEMININE on Ciswomen	0.495	0.508	0.432	0.374	0.452
H_0: MASCULINE on Ciswomen	0.495	0.374	0.368	0.387	0.426
H_0: FEMININE on Transmen	0.851	0.953	0.928	0.821	0.832
H_0: MASCULINE on Transmen	0.891	0.761	0.922	0.779	0.828
H_0: FEMININE on Transwomen	0.669	0.541	0.274	0.310	0.386
H_0: MASCULINE on Transwomen	0.549	0.449	0.627	0.724	0.591

Note: Competition is a binary variable equal to 1 if the participant enters the tournament in Part 4 and 0 otherwise. Delta perf. is the difference in performance between Part 3 (tournament) and Part 2 (piece-rate). Belief tournament is the participants' belief of their performance rank within their group in Part 3, where the value 1 represents the rank with the highest performance. Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious cisman, who earns more than 20K GBP, and lives in continental Europe. In the last column from the right, the baseline is a cisman. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Rows starting with H_0 report the p-values of a joint coefficient test that the coefficients' sum equals 0. For example, "H_0: FEMININE on Ciswomen" tests the effect of the treatment (FEMININE) on the subject group (Ciswomen).

C.6 Payoffs (Part 4)

C.6.1 Regressions

Table C.17. OLS regression for NEUTRAL. Dependent variable: Payoff in Part 4.

	(1)	(2)	(3)
Ciswomen	-0.156	1.315	-0.267
	(1.183)	(1.326)	(1.177)
Transmen	0.951	2.316	3.377
	(1.256)	(1.358)	(2.170)
Transwomen	-1.071	-0.560	1.571
	(1.244)	(1.278)	(2.215)
Age		-0.040	
		(0.047)	
Height		0.113 **	
		(0.041)	
Student status		3.479 ***	
		(1.017)	
Income: < 20,000 GBP		0.402	
		(0.922)	
Religion: Religious		-2.336 *	
		(0.953)	
Religion: Not say		-3.673 **	
		(1.247)	
Residence: US		1.421	
		(1.212)	
Residence: UK		1.468	
		(1.176)	
Residence: Other		-1.269	
		(1.576)	
TCS			1.036
			(0.634)
STT			-0.099
			(0.153)
Const.	5.389 ***	-15.595 *	0.895
	(0.851)	(7.668)	(2.853)
N	259	259	259
Adj. R2	-0.003	0.073	-0.002
H_0: Sex	0.289	0.056	0.366
H_0: Gender	0.215	0.381	0.234

Note: Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious cisman, who earns more than 20K GBP, and lives in continental Europe. In the last column from the right, the baseline is a cisman. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. H_0: Sex reports the p-values of a joint coefficient test comparing Male (Cismen and Transwomen) with Female (Ciswomen and Transmen). H_0: Gender reports the p-values of a joint coefficient test comparing Masculine (Cismen and Transmen) with Feminine (Ciswomen and Transwomen).

Table C.18. OLS regression for all treatments. Dependent variable: Payoff in Part 4.

	(1)	(2)	(3)
Ciswomen	-0.156	0.922	-0.225
	(1.183)	(1.245)	(1.178)
Transmen	0.951	1.785	1.993
	(1.256)	(1.279)	(1.614)
Transwomen	-1.071	-0.976	0.193
	(1.244)	(1.250)	(1.643)
Treatment: FEMININE	2.660	2.622	2.632
	(1.516)	(1.483)	(1.525)
Treatment: MASCULINE	0.442	0.354	0.435
	(1.331)	(1.343)	(1.328)
FEMININE x Ciswomen	-2.293	-2.286	-2.207
	(1.925)	(1.892)	(1.931)
MASCULINE x Ciswomen	-1.096	-1.243	-1.003
	(1.700)	(1.717)	(1.692)
FEMININE x Transmen	-4.501 *	-4.672 *	-4.383 *
	(1.905)	(1.868)	(1.909)
MASCULINE x Transmen	-0.818	-0.744	-0.781
	(1.859)	(1.850)	(1.849)
FEMININE x Transwomen	-1.689	-1.536	-1.664
	(2.111)	(2.116)	(2.108)
MASCULINE x Transwomen	-0.075	-0.088	-0.054
	(1.813)	(1.770)	(1.827)
Const.	5.389 ***	-8.821	2.454
	(0.851)	(5.310)	(1.658)
Controls (Age, Height, Student status, Income, Religion, Residence)	–	Yes	–
Controls (TCS, STT)	–	–	Yes
N	780	780	780
Adj. R2	0.003	0.034	0.005
H_0: FEMININE on Cismen	0.080	0.077	0.085
H_0: MASCULINE on Cismen	0.740	0.792	0.743
H_0: FEMININE on Ciswomen	0.757	0.773	0.719
H_0: MASCULINE on Ciswomen	0.536	0.411	0.589
H_0: FEMININE on Transmen	0.111	0.070	0.129
H_0: MASCULINE on Transmen	0.772	0.760	0.788
H_0: FEMININE on Transwomen	0.509	0.472	0.514
H_0: MASCULINE on Transwomen	0.766	0.822	0.762

Note: Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious cisman, who earns more than 20K GBP, and lives in continental Europe. In the last column from the right, the baseline is a cisman. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Rows starting with H_0 report the p-values of a joint coefficient test that the coefficients' sum equals 0. For example, "H_0: FEMININE on Ciswomen" tests the effect of the treatment (FEMININE) on the subject group (Ciswomen).

C.7 Risk (Part 5)

C.7.1 Bar graphs

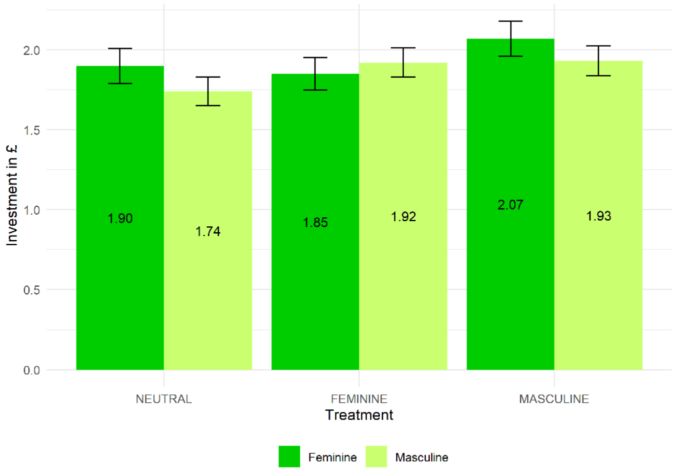


Figure C.4. Investment into the risky lottery in Part 5 by treatments and gender ($n = 780$).

Note: The bars show the average investment rate, and the error bars represent the standard error of the mean.

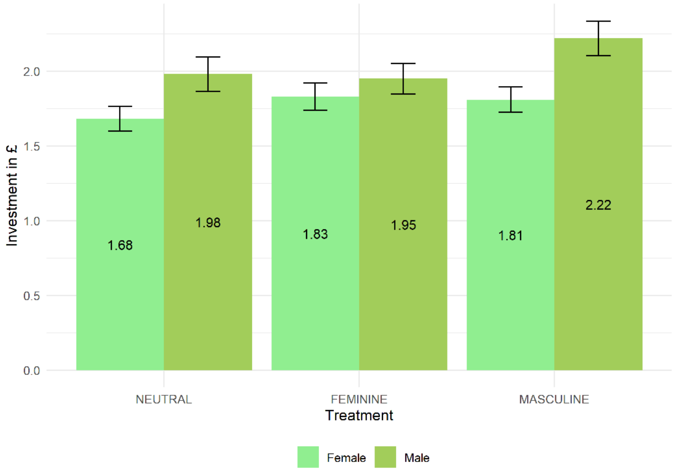


Figure C.5. Investment into the risky lottery in Part 5 by treatments and sex ($n = 780$).

Note: The bars show the average investment rate, and the error bars represent the standard error of the mean.

C.7.2 Non-parametric tests

Table C.19. Investment into the risky lottery across treatments and subject groups.

Panel A: Risk across treatments

	Treatment			
Subject groups	NEUTRAL	FEMININE	MASCULINE	p-value
Cismen	1.814	2.021	2.208	0.119
Ciswomen	1.690	1.852	1.972	0.446
Transmen	1.673	1.816	1.655	0.660
Transwomen	2.244	1.840	2.227	0.357

Panel B: Risk across subject groups

	Subject groups				
Treatment	Cismen	Ciswomen	Transmen	Transwomen	p-value
NEUTRAL	1.814	1.690	1.673	2.244	0.194
FEMININE	2.021	1.852	1.816	1.840	0.715
MASCULINE	2.208	1.972	1.655	2.227	0.030

Panel C: Risk across groups within NEUTRAL

	Group 1		Group 2		
	Subjects		Subjects		p-value
Case 1	Cisgender	1.753	Transgender	1.890	0.461
Case 2	Cismen	1.814	Ciswomen	1.690	0.704
Case 3	Transmen	1.673	Transwomen	2.244	0.048
Case 4	Female	1.681	Male	1.977	0.130
Case 5	Feminine	1.902	Masculine	1.743	0.355

Panel D: Risk in NEUTRAL compared to the other treatments

	NEUTRAL	FEMININE		MASCULINE	
Subject groups			p-value		p-value
Cismen	1.814	2.021	0.262	2.208	0.038
Ciswomen	1.690	1.852	0.550	1.972	0.208
Transmen	1.673	1.816	0.479	1.655	0.881
Transwomen	2.244	1.840	0.206	2.227	0.927

Note: The column p-value reports the results of the tests performed column-wise. For continuous variables, we conducted Mann-Whitney U tests for two groups and Kruskal-Wallis tests for more than two groups.

C.7.3 Regressions

Table C.20. OLS regression for NEUTRAL. Dependent variable: Risk.

	(1)	(2)	(3)
Ciswomen	-0.124	-0.164	-0.109
	(0.179)	(0.218)	(0.177)
Transmen	-0.141	-0.221	-0.295
	(0.177)	(0.212)	(0.396)
Transwomen	0.430	0.316	0.216
	(0.245)	(0.258)	(0.412)
Age		-0.008	
		(0.010)	
Height		-0.001	
		(0.008)	
Student status		0.064	
		(0.161)	
Income: < 20,000 GBP		-0.125	
		(0.162)	
Religion: Religious		-0.045	
		(0.160)	
Religion: Not say		-0.502	
		(0.414)	
Residence: US		0.052	
		(0.202)	
Residence: UK		0.145	
		(0.194)	
Residence: Other		0.618 *	
		(0.309)	
TCS			-0.143
			(0.105)
STT			-0.009
			(0.029)
Const.	1.814 ***	2.273	2.463 ***
	(0.133)	(1.565)	(0.499)
N	259	259	259
Adj. R2	0.022	0.018	0.024
H_0: Sex	0.020	0.042	0.037
H_0: Gender	0.132	0.228	0.183

Note: Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious cisman, who earns more than 20K GBP, and lives in continental Europe. In the last column from the right, the baseline is a cisman. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. H_0: Sex reports the p-values of a joint coefficient test comparing Male (Cismen and Transwomen) with Female (Ciswomen and Transmen). H_0: Gender reports the p-values of a joint coefficient test comparing Masculine (Cismen and Transmen) with Feminine (Ciswomen and Transwomen).

Table C.21. OLS regression for all treatments. Dependent variable: Risk.

	(1)	(2)	(3)
Ciswomen	-0.124	-0.123	-0.124
	(0.179)	(0.195)	(0.179)
Transmen	-0.141	-0.154	-0.273
	(0.177)	(0.189)	(0.286)
Transwomen	0.430	0.387	0.325
	(0.245)	(0.248)	(0.324)
Treatment: FEMININE	0.207	0.210	0.218
	(0.192)	(0.193)	(0.192)
Treatment: MASCULINE	0.394 *	0.386 *	0.393 *
	(0.191)	(0.193)	(0.191)
FEMININE x Ciswomen	-0.045	-0.048	-0.052
	(0.263)	(0.263)	(0.263)
MASCULINE x Ciswomen	-0.112	-0.118	-0.106
	(0.256)	(0.259)	(0.255)
FEMININE x Transmen	-0.064	-0.074	-0.073
	(0.255)	(0.257)	(0.255)
MASCULINE x Transmen	-0.412	-0.409	-0.410
	(0.252)	(0.253)	(0.252)
FEMININE x Transwomen	-0.611	-0.593	-0.585
	(0.319)	(0.320)	(0.322)
MASCULINE x Transwomen	-0.412	-0.380	-0.403
	(0.348)	(0.350)	(0.349)
Const.	1.814 ***	1.605	1.762 ***
	(0.133)	(0.913)	(0.308)
Controls (Age, Height, Student status, Income, Religion, Residence)	–	Yes	–
Controls (TCS, STT)	–	–	Yes
N	780	780	780
Adj. R2	0.018	0.013	0.017
H_0: FEMININE on Cismen	0.282	0.278	0.257
H_0: MASCULINE on Cismen	0.039	0.046	0.040
H_0: FEMININE on Ciswomen	0.368	0.370	0.356
H_0: MASCULINE on Ciswomen	0.097	0.125	0.092
H_0: FEMININE on Transmen	0.395	0.426	0.388
H_0: MASCULINE on Transmen	0.915	0.889	0.918
H_0: FEMININE on Transwomen	0.113	0.132	0.156
H_0: MASCULINE on Transwomen	0.951	0.984	0.975

Note: Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious cisman, who earns more than 20K GBP, and lives in continental Europe. In the last column from the right, the baseline is a cisman. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Rows starting with H_0 report the p-values of a joint coefficient test that the coefficients' sum equals 0. For example, "H_0: FEMININE on Ciswomen" tests the effect of the treatment (FEMININE) on the subject group (Ciswomen).

C.8 Altruism (Part 6)

C.8.1 Bar graphs

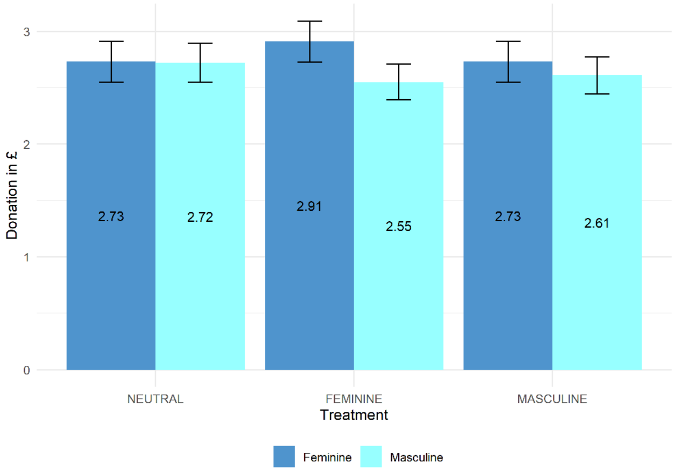


Figure C.6. Donation in Part 6 by treatments and gender ($n = 780$).

Note: The average donations are indicated by the bars and the error bars represent the standard error of the mean.

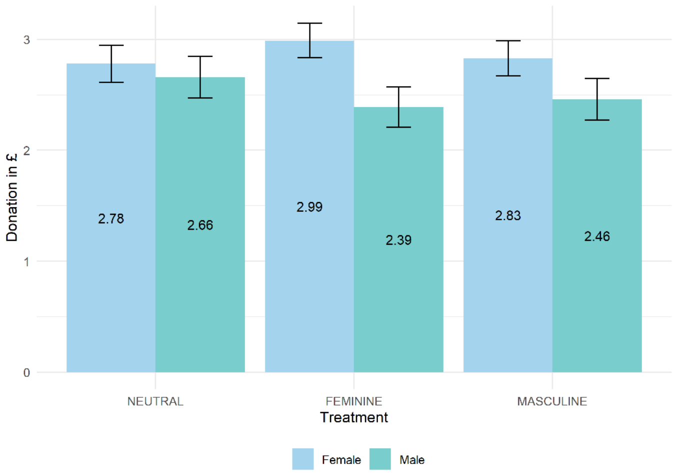


Figure C.7. Donation in Part 6 by treatments and sex ($n = 780$).

Note: The average donations are indicated by the bars and the error bars represent the standard error of the mean.

C.8.2 Non-parametric tests

Table C.22. Donations across treatments and subject groups.

Panel A: Donations across treatments

	Treatment			
Subject groups	NEUTRAL	FEMININE	MASCULINE	p-value
Cismen	2.685	2.265	2.423	0.582
Ciswomen	2.803	3.161	2.864	0.478
Transmen	2.762	2.822	2.792	0.999
Transwomen	2.615	2.556	2.525	0.999

Panel B: Donations across subject groups

	Subject groups				
Treatment	Cismen	Ciswomen	Transmen	Transwomen	p-value
NEUTRAL	2.685	2.803	2.762	2.615	0.933
FEMININE	2.265	3.161	2.822	2.556	0.073
MASCULINE	2.423	2.864	2.792	2.525	0.540

Panel C: Donations across groups within NEUTRAL

	Group 1		Group 2		
	Subjects		Subjects		p-value
Case 1	Cisgender	2.743	Transgender	2.706	0.871
Case 2	Cismen	2.685	Ciswomen	2.803	0.759
Case 3	Transmen	2.762	Transwomen	2.615	0.583
Case 4	Female	2.782	Male	2.658	0.564
Case 5	Feminine	2.731	Masculine	2.723	0.914

Panel D: Donations in NEUTRAL compared to the other treatments

	NEUTRAL	FEMININE		MASCULINE	
Subject groups			p-value		p-value
Cismen	2.685	2.265	0.322	2.423	0.454
Ciswomen	2.803	3.161	0.260	2.864	0.863
Transmen	2.762	2.822	0.982	2.792	0.977
Transwomen	2.615	2.556	0.948	2.525	0.987

Note: The column p-value reports the results of the tests performed column-wise. For continuous variables, we conducted Mann-Whitney U tests for two groups and Kruskal-Wallis tests for more than two groups.

C.8.3 Regressions

Table C.23. OLS regression for NEUTRAL. Dependent variable: Donations.

	(1)	(2)	(3)
Ciswomen	0.118	0.188	0.128
	(0.335)	(0.403)	(0.337)
Transmen	0.077	0.239	-0.162
	(0.343)	(0.387)	(0.646)
Transwomen	-0.070	0.103	-0.324
	(0.378)	(0.409)	(0.647)
Age		0.040 *	
		(0.016)	
Height		0.002	
		(0.013)	
Student status		0.713 **	
		(0.274)	
Income: < 20,000 GBP		0.025	
		(0.279)	
Religion: Religious		0.800 **	
		(0.274)	
Religion: Not say		0.701	
		(0.736)	
Residence: US		-0.789 *	
		(0.324)	
Residence: UK		-0.470	
		(0.335)	
Residence: Other		-0.391	
		(0.538)	
TCS			-0.091
			(0.199)
STT			0.012
			(0.046)
Const.	2.685 ***	1.123	3.077 ***
	(0.244)	(2.413)	(0.909)
N	259	259	259
Adj. R2	-0.011	0.044	-0.018
H_0: Sex	0.600	0.600	0.580
H_0: Gender	0.955	0.920	0.948

Note: Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious cisman, who earns more than 20K GBP, and lives in continental Europe. In the last column from the right, the baseline is a cisman. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. H_0: Sex reports the p-values of a joint coefficient test comparing Male (Cismen and Transwomen) with Female (Ciswomen and Transmen). H_0: Gender reports the p-values of a joint coefficient test comparing Masculine (Cismen and Transmen) with Feminine (Ciswomen and Transwomen).

Table C.24. OLS regression for all treatments. Dependent variable: Donations.

	(1)	(2)	(3)
Ciswomen	0.118	0.175	0.128
	(0.335)	(0.358)	(0.336)
Transmen	0.077	0.208	-0.036
	(0.343)	(0.351)	(0.465)
Transwomen	-0.070	-0.004	-0.222
	(0.378)	(0.386)	(0.477)
Treatment: FEMININE	-0.420	-0.377	-0.419
	(0.338)	(0.343)	(0.338)
Treatment: MASCULINE	-0.262	-0.229	-0.261
	(0.340)	(0.346)	(0.340)
FEMININE x Ciswomen	0.778	0.745	0.767
	(0.468)	(0.471)	(0.469)
MASCULINE x Ciswomen	0.323	0.355	0.308
	(0.468)	(0.470)	(0.468)
FEMININE x Transmen	0.480	0.410	0.465
	(0.468)	(0.458)	(0.468)
MASCULINE x Transmen	0.292	0.270	0.287
	(0.474)	(0.466)	(0.474)
FEMININE x Transwomen	0.361	0.444	0.350
	(0.530)	(0.533)	(0.529)
MASCULINE x Transwomen	0.172	0.191	0.166
	(0.543)	(0.547)	(0.545)
Const.	2.685 ***	1.806	3.117 ***
	(0.244)	(1.435)	(0.516)
Controls (Age, Height, Student status, Income, Religion, Residence)	–	Yes	–
Controls (TCS, STT)	–	–	Yes
N	780	780	780
Adj. R2	-0.000	0.023	-0.002
H_0: FEMININE on Cismen	0.214	0.272	0.216
H_0: MASCULINE on Cismen	0.441	0.509	0.444
H_0: FEMININE on Ciswomen	0.270	0.256	0.282
H_0: MASCULINE on Ciswomen	0.848	0.693	0.882
H_0: FEMININE on Transmen	0.854	0.914	0.886
H_0: MASCULINE on Transmen	0.926	0.895	0.937
H_0: FEMININE on Transwomen	0.885	0.870	0.867
H_0: MASCULINE on Transwomen	0.832	0.930	0.825

Note: Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious cisman, who earns more than 20K GBP, and lives in continental Europe. In the last column from the right, the baseline is a cisman. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Rows starting with H_0 report the p-values of a joint coefficient test that the coefficients' sum equals 0. For example, "H_0: FEMININE on Ciswomen" tests the effect of the treatment (FEMININE) on the subject group (Ciswomen).

C.9 Continuous gender measure (BEM)

C.9.1 Competitiveness

Table C.25. Probit regression for NEUTRAL. Dependent variable: Competition. Gender is measured on a continuous scale.

	(1)	(2)	(3)	(4)	(5)
BEM score: Feminine	0.010	0.014	0.016	0.017	0.016
	(0.011)	(0.011)	(0.012)	(0.013)	(0.012)
BEM score: Masculine	0.022	0.019	0.015	0.016	0.015
	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)
Perf. tournament		0.004	-0.064 *	-0.057	-0.063 *
		(0.024)	(0.028)	(0.029)	(0.029)
Delta perf.		0.121 ***	0.097 *	0.108 **	0.097 *
		(0.034)	(0.038)	(0.038)	(0.039)
Belief tournament			-0.500 ***	-0.484 ***	-0.501 ***
			(0.123)	(0.126)	(0.122)
Risk			0.129	0.131	0.129
			(0.077)	(0.079)	(0.077)
Age				0.023	
				(0.012)	
Height				-0.000	
				(0.009)	
Student status				-0.114	
				(0.204)	
Income: < 20,000 GBP				0.464 *	
				(0.231)	
Religion: Religious				-0.001	
				(0.212)	
Religion: Not say				0.227	
				(0.471)	
Residence: US				0.094	
				(0.255)	
Residence: UK				0.391	
				(0.260)	
Residence: Other				0.495	
				(0.380)	
TCS					-0.012
					(0.096)
STT					-0.011
					(0.022)
Const.	-1.677 **	-1.847 **	-0.315	-1.493	-0.224
	(0.585)	(0.651)	(0.803)	(1.759)	(0.889)
N	259	259	259	259	259
Pseudo R2 (McFadden)	0.018	0.074	0.139	0.175	0.140

Note: Competition is a binary variable equal to 1 if the participant enters the tournament in Part 4 and 0 otherwise. Delta perf. is the difference in performance between Part 3 (tournament) and Part 2 (piece-rate). Belief tournament is the participants' belief of their performance rank within their group in Part 3, where the value 1 represents the rank with the highest performance. Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious participant, who earns more than 20K GBP, and lives in continental Europe. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

C.9.2 Risk

Table C.26. OLS regression for NEUTRAL. Dependent variable: Risk. Gender is measured on a continuous scale.

	(1)	(2)	(3)
BEM score: Feminine	-0.009	-0.008	-0.009
	(0.009)	(0.009)	(0.009)
BEM score: Masculine	0.006	0.006	0.010
	(0.009)	(0.010)	(0.009)
Age		-0.008	
		(0.009)	
Height		0.005	
		(0.007)	
Student status		0.045	
		(0.160)	
Income: < 20,000 GBP		-0.091	
		(0.161)	
Religion: Religious		-0.067	
		(0.167)	
Religion: Not say		-0.518	
		(0.419)	
Residence: US		0.091	
		(0.204)	
Residence: UK		0.167	
		(0.190)	
Residence: Other		0.769 *	
		(0.302)	
TCS			-0.173 *
			(0.078)
STT			-0.019
			(0.018)
Const.	2.011 ***	1.262	2.577 ***
	(0.440)	(1.339)	(0.568)
N	259	259	259
Adj. R2	-0.002	0.004	0.012

Note: Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious participant, who earns more than 20K GBP, and lives in continental Europe. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

C.9.3 Altruism

Table C.27. OLS regression for NEUTRAL. Dependent variable: Donations. Gender is measured on a continuous scale.

	(1)	(2)	(3)
BEM score: Feminine	0.023	0.018	0.023
	(0.015)	(0.015)	(0.016)
BEM score: Masculine	0.002	-0.014	0.003
	(0.016)	(0.016)	(0.016)
Age		0.040 *	
		(0.017)	
Height		-0.002	
		(0.010)	
Student status		0.710 *	
		(0.278)	
Income: < 20,000 GBP		0.051	
		(0.275)	
Religion: Religious		0.797 **	
		(0.279)	
Religion: Not say		0.608	
		(0.722)	
Residence: US		-0.831 **	
		(0.309)	
Residence: UK		-0.480	
		(0.330)	
Residence: Other		-0.401	
		(0.499)	
TCS			-0.031
			(0.133)
STT			-0.002
			(0.031)
Const.	1.713 *	1.573	1.808
	(0.774)	(1.963)	(0.919)
N	259	259	259
Adj. R2	0.001	0.054	-0.007

Note: Standard errors in parentheses are heteroskedasticity robust. In the second last column from the right, the baseline is a non-student, non-religious participant, who earns more than 20K GBP, and lives in continental Europe. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

C.10 Detailed literature summary

C.10.1 Risk

Risk-taking is considered a fundamental determinant of individual behavior in different domains like health (Anderson and Mellor, 2008; Barsky et al., 1997), stock market participation (Almenberg and Dreber, 2015), saving decisions (Sutter et al., 2013), occupational and self-employment choices (Bonin et al., 2007), personal and household finance (Bucciol and Miniaci, 2011; Guiso and Paiella, 2008), education (Von Gaudecker et al., 2011) and environmental decision-making (Gollier, 2001). The literature reports strong evidence for women preferring to take less risk compared to men (Charness and Gneezy, 2012). This difference in risk-taking is robust when using different experimental methods to measure risk like lotteries (Holt and Laury, 2002), investment games (Gneezy and Potters, 1997) or card games (Czibor et al., 2019). It is also reported for subjects varying from children (Cárdenas et al., 2015), to students (Croson and Gneezy, 2009), to non-students (Hardies et al., 2013). Moreover, the difference is not influenced by conducting the experiment in the lab or in other environments like on online platforms (Hardies et al., 2013).

Several studies analyze gender differences in risk preferences for sub-populations of managers (Adams and Funk, 2012; Atkinson et al., 2003; Croson and Gneezy, 2009) and find that females are similar or even less risk-averse than men. The reasons could be a selection or social learning and adaptive behavior to the job demands. To disentangle these different factors, Drupp et al. (2020) use an online experiment with scientists. They vary the salience of either the private or the professional identity of the subjects. They report that priming the professional identity reduces the gender gap in risk-taking. Besides, the gender gap decreases with increasing age as female senior scientists choose riskier options in the treatment where the profession is made salient.

Also, attempts to explore the connection between biological factors and risk-taking are taken for the domain of risky behavior. First, studies are exploring the causal effect of hormones on behavior.[33] For example, Zethraeus et al. (2009) test for administered testosterone or estrogen affecting women's risk-taking. No effect of either testosterone or estrogen on risk-taking could be detected. In line, the study by Boksem et al. (2013) and Buskens et al. (2016) find no effect of testosterone on risk aversion. Ranehill et al. (2018) take a comparative approach and administer an oral contraceptive or not. Again, no connection between hormones and behavior is reported. Second, studies test for the correlation between the variation in risk-taking and genes. On the one hand, for example, Anderson et al. (2015) find no relationship between the dopamine and the serotonin gene and risk-taking. On the other hand, studies using, for example, the twin methodology and *genome-wide association techniques* (GWAS) report genetic foundations for the willingness to take risk (Cesarini et al., 2009, 2010, 2012). Third, a recent study by Keaveney et al.

[33]For our literature review, we summarize only studies that concentrate on pharmacological testosterone administration with double-blind placebo-controlled designs, which allows us to interpret results causally.

(2020) showed that the intake of a small dose of Acetaminophen, a very popular pain killer, increases risk-taking.

Several researchers prime subjects and study the effect on risk-taking (Erb et al., 2002; Gilad and Kliger, 2008; Guiso et al., 2018; König-Kersting and Trautmann, 2018; Newell and Shaw, 2017). The study closest to our research is Benjamin et al. (2010) which finds that making the subject's gender salient with a short questionnaire does not impact risk preferences. Also, Meier-Pesti and Penz (2008) report an effect of gender priming through questions and stereotypical pictures only on male risk preferences. Cohn et al. (2017) prime financial professionals with their professional salience, which leads to a decrease in risk-taking in a high stakes investment game. With a similar subject pool, Cohn et al. (2015) find that individuals primed with a bust scenario are more risk-averse compared to those primed with a boom scenario. Alempaki et al. (2019) test the robustness of the results of Cohn et al. (2015) with an Amazon Mechanical Turk subject pool. They report no evidence of priming influencing risk-taking. Callen et al. (2014) primed individuals who were exposed to violence by asking them to either recall happy, fearful or neutral moments. They find that remembering frightening experiences leads to a higher preference for certainty.

The only related study we are aware of that investigates the risk-taking behavior of LGBTQ+ individuals is Buser et al. (2018a). It analyzes risk preferences by asking the subjects about their risk perception (survey question). It finds no significant differences between homosexual and heterosexual men and homosexual and heterosexual women.

C.10.2 Altruism

To what extend someone is pro-social, i.e., altruistic, is argued to explain behavior in the labor market, how individuals vote, if they take up volunteer work or not, and how willing someone is to give to a charity (Bilén et al., 2021). Altruistic behavior is typically measured with a dictator game, where participants are asked how much they want to transfer to an anonymous other participant (Forsythe et al., 1994; Kahneman et al., 1986), or how much they wish to donate to a charity (Eckel and Grossman, 1996). It is a robust finding that participants in experiments transfer quite a substantial part of their endowment in dictator games, thus act altruisticly (Carpenter et al., 2008). The literature reports mixed findings on the external validity of those experiments. One strand of the literature finds that individuals behave in donation experiments similar as in naturally occurring decision situations on charitable giving (Benz and Meier, 2008; Franzen and Pointner, 2013). Other research contradicts these findings, as recently summarized by Galizzi and Navarro-Martinez (2019).

Concerning the level of altruism exhibited by men and women, a wide range of studies shows that women are generally more generous in dictator games. See, e.g., Bilén et al. (2021) for an up-to-date meta-analysis of the existing literature on gender differences in charitable giving. These authors report that the magnitude of the gender differences in altruism is sensible to the experimental context. For example, the difference is more prominent if the dictator decides to donate to a charity than giving to an anonymous recipient. However, the difference is more minor if the dictator chooses between giving all or nothing compared to deciding on a continuous scale.

Turning to studies attempting to link hormones to altruism causally, Buskens et al. (2016) and Zak et al. (2009) found no impact of administered testosterone on dictators' giving. Zethraeus et al. (2009) used another approach and administered testosterone, estrogen, or a placebo to the experimental participants. Again, no connection between either hormone or altruism is reported. Moreover, administering an oral contraceptive containing synthetic progesterone as the main ingredient suggests no hormonal impact on altruism levels. However, there is evidence that the underlying genes influence altruism. See for example Reuter et al. (2011), who used twins for their study.

Multiple studies explore if different priming influences altruistic behavior. For example, subsequent donations are affected by religious primes (Ahmed and Salas, 2011; Benjamin et al., 2016; McKay et al., 2011; Shariff and Norenzayan, 2007), by reminding subjects of secular, moral institutions (Shariff and Norenzayan, 2007), and by priming with subtle cues of observability (Bateson et al., 2006; Haley and Fessler, 2005; Rigdon et al., 2009). Boschini et al. (2018) report an increased gender gap in altruism when making gender more salient by requiring participants to specify their gender before the dictator game and informing them about the gender of the recipient. Again, we have found no published studies of altruism of LGBTQ+ individuals in economics.

C.11 Additional information

C.11.1 Study sample

We recruited a total of 798 participants. Please note that due to a technical problem on how the participant's performance was shown to them on their screen, we exclude $n = 3$ cisgender and $n = 6$ transgender observations. We tested with our debriefing questionnaire whether the participants had an idea about the aim of the study, the study topic, etc. Eight cisgender and one transgender participant(s) wrote to think that s/he were primed. These $n = 9$ observations are also excluded from our analysis. Thus, the final number of subjects by subject groups and treatment is 780 as summarized in Table C.28.

To have comparable transgender and cisgender observations, we first collected the major part of the transgender observations, including their main demographic characteristics (age, student status, education, income, religious affiliation, and residence) of the transgender participants. We then used Prolific's sorting tool to recruit a similar cisgender sample based on those criteria.

Table C.28. Distribution of subject groups across treatments.

	Treatments			
	NEUTRAL	FEMININE	MASCULINE	Total
Cismen	72	71	71	214
Ciswomen	71	70	70	211
Transmen	72	72	71	215
Transwomen	44	50	46	140
Total	259	263	258	780

Note: The table summarizes the number of participants of the four subject groups (cismen, ciswomen, transmen, transwomen) in the three treatments (NEUTRAL, FEMININE, MASCULINE).

C.11.2 Data sets

gender_data.csv This is the main data set. The file contains n=798 observations and 103 variables. Details on the variables can be found in the second data set (codebook.csv).

codebook.csv This file provides the details on the variables of the main data set. Each row includes the explanations for one of the 103 variables. Additionally, the third column summarizes the response options the subjects had.

The collected data and additional material is available at OSF using `Link` or Online (2022g).

C.11.3 Instructions

The following pages contain screenshots of the online study conducted on the platform Prolific. Please note that one participant was randomly allocated to just one treatment. Thus, one participant saw one out of the three different treatment pages. In addition, depending on the choice made in Part 4, the system showed one of two option pages. A blue headline marks the varying screens. All other pages were identical.

Welcome!

Please read the following.

Dear participant,
The following will provide you with information about the experiment that will help you decide whether you wish to participate. The study received two certificates of good standing (ethical approvals). If you agree to participate, please be aware that you are free to withdraw at any point throughout the study. All data collected for this scientific study will remain confidential and anonymous. If you have any further questions concerning this study, please feel free to contact us via phone or email:

Dr. Silvio Städter at silvio.staedter@ur.de or +49 941 9433259, or the team at econ.study.research@gmail.com.

Please indicate that by clicking the following box on the space below, you understand your rights and agree to participate in the experiment. You can revoke the consent to the collection and processing of the data at any time by just closing the internet browser via which you participate. After your revocation, no further data will be collected. However, the data collected up to the point of cancellation can continue to be used in this study. Your participation is solicited yet strictly voluntary. All information will be kept confidential, and your name will not be associated with any research findings.

☐ *I understand my rights and agree to participate in the experiment.*

Continue

Welcome!

Thank you very much for participating in our study!

This is a project of researchers from the University of Exeter (United Kingdom), the Chulalongkorn University (Thailand), and the University of Regensburg (Germany). The study is conducted by Prof. Brit Grosskopf, Chanalak Chaisrilak, Dr. Helena Fornwagner, Alexander Lauf, Vanessa Schöller, and Dr. Silvio Städter.

The study received two certificates of good standing (ethical approvals).

The study consists of **six parts** and a **questionnaire**.

Depending on the decisions you take in the different parts, you can earn money.

Please note that all decisions you take, as well as all data that is collected with the survey is anonymous and only used for this study.

Before each part starts, you will get detailed instructions, what you have to do and how the decisions you take influence your payment.

In each part, you can earn money. At the end of the study, you can see how much you earned in each part. We will then randomly select one out of the six parts and pay you the money you have earned in this part.

For completing the study you will receive a compensation of **£3.75** for sure.

So your total payment will be your earnings in one randomly chosen part plus the compensation for completing the study. At the end of the study, you will receive your personal unique **8-digit Code** as proof that you have completed the study.

If you need any further details or have problems when conducting the study, please contact Dr. Silvio Städter at silvio.staedter@ur.de or +49 941 9433259, or the team at econ.study.research@gmail.com.

Please enter your personal Code in Prolific. Then, you get **£3.75** and the additional money you earned in the randomly selected part via Prolific.

Thank you again for taking your time to participate,
Prof. Brit Grosskopf,
Chanalak Chaisrilaky, MSc,
Dr. Helena Fornwagner,
Alexander Lauf, MSc,
Vanessa Schöller, MSc, and
Dr. Silvio Städter.

Start the Study

Part 1: Instructions

In this game you have to solve a **Word Search Puzzle**.

On the next page, you will see a table with **10** rows and **10** columns, a word puzzle. Make sure you can see the entire table.

Within this table **8 words** are hidden. They are either **vertically**, **horizontally** or **diagonally** placed in the table.

These 8 words are revealed on the next page, just above the table. Your task is to find these 8 hidden words and mark them. To mark a word, you have to keep the left mouse button down and go over the word's letters.

Marked words are highlighted in blue. If you mark a word, it will be crossed out in the list of words. You have to mark one word in one go.

If this part is randomly selected for payment at the end of the study and you have found **all 8 words**, you get **£5.00**.
If you didn't find all 8 words, you get nothing in this part.

You cannot go back to this page and please do not reload the page.

You have **120 seconds** to find the 8 words.

If you are ready, click on the button 'Start this part' below.

Start this part

Treatment: NEUTRAL

Part 1: Word Search

Time left to complete this page: **1:57**

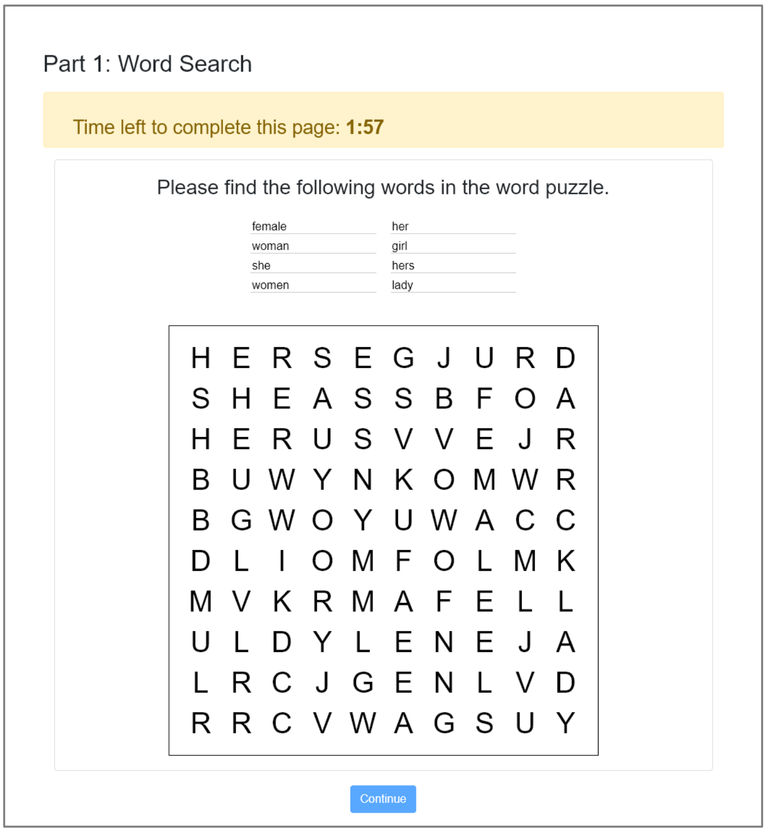
Please find the following words in the word puzzle.

female	her
woman	girl
she	hers
women	lady

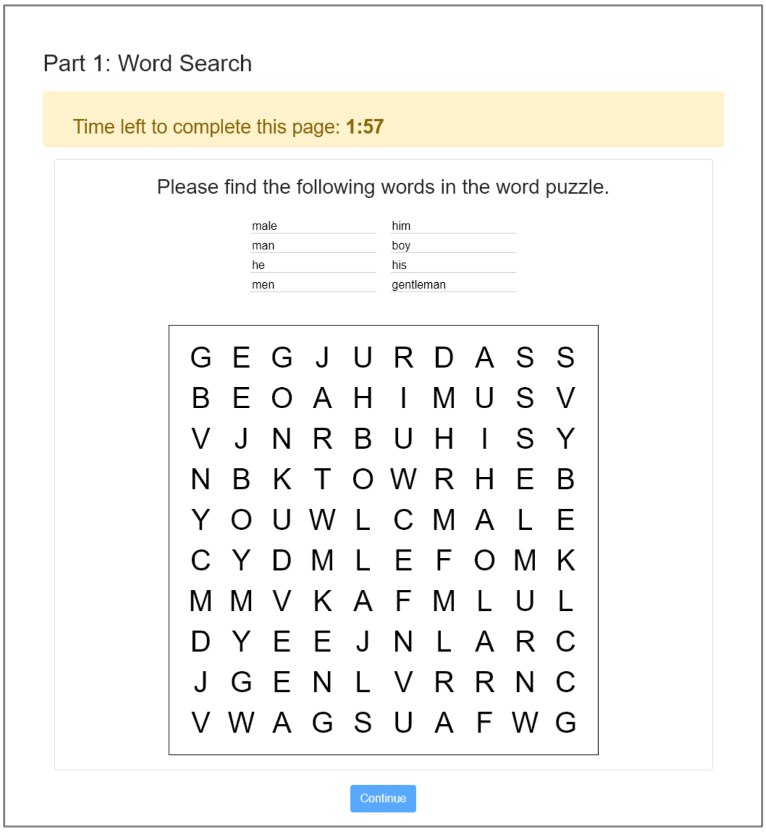
H	E	R	S	E	G	J	U	R	D
S	H	E	A	S	S	B	F	O	A
H	E	R	U	S	V	V	E	J	R
B	U	W	Y	N	K	O	M	W	R
B	G	W	O	Y	U	W	A	C	C
D	L	I	O	M	F	O	L	M	K
M	V	K	R	M	A	F	E	L	L
U	L	D	Y	L	E	N	E	J	A
L	R	C	J	G	E	N	L	V	D
R	R	C	V	W	A	G	S	U	Y

Continue

Treatment: FEMININE



Treatment: MASCULINE



Part 1: Summary

You have finished this part.

You found **8 of 8** words.

Continue to next part

Part 2, Part 3, and Part 4

Please complete Part 2, Part 3 and Part 4 now.

In each part, you have to perform the **same task**. In this task, you have to solve as many short math quizzes as you can within 120 seconds in each part. The number of correctly solved quizzes determines your payment in each part.

Before you can start to earn money in each part, you get detailed instructions on how your payment is determined in the following part.

After the instructions for the first part, you get to practice the task for 90 seconds. The quizzes solved for practice purposes are not payoff relevant.

Continue

Part 2: Instructions

Instructions: Math Quizzes

In this task, you have to solve as many quizzes as possible within 120 seconds. For each correctly solved quiz you can earn money as explained under '*Payment*' after the practice round.

Each quiz consists of 9 two-digit numbers. You can see an example of a quiz below.

Quiz 1

You solved 0 of 0 correctly.

Please select 2 numbers that add up to 100.

10	15	56
87	75	25
20	31	14

Submit

In each quiz, you have to select two numbers that add up to 100 by clicking on them. If you have selected two numbers, you can submit your selection by clicking on the button 'Submit'.

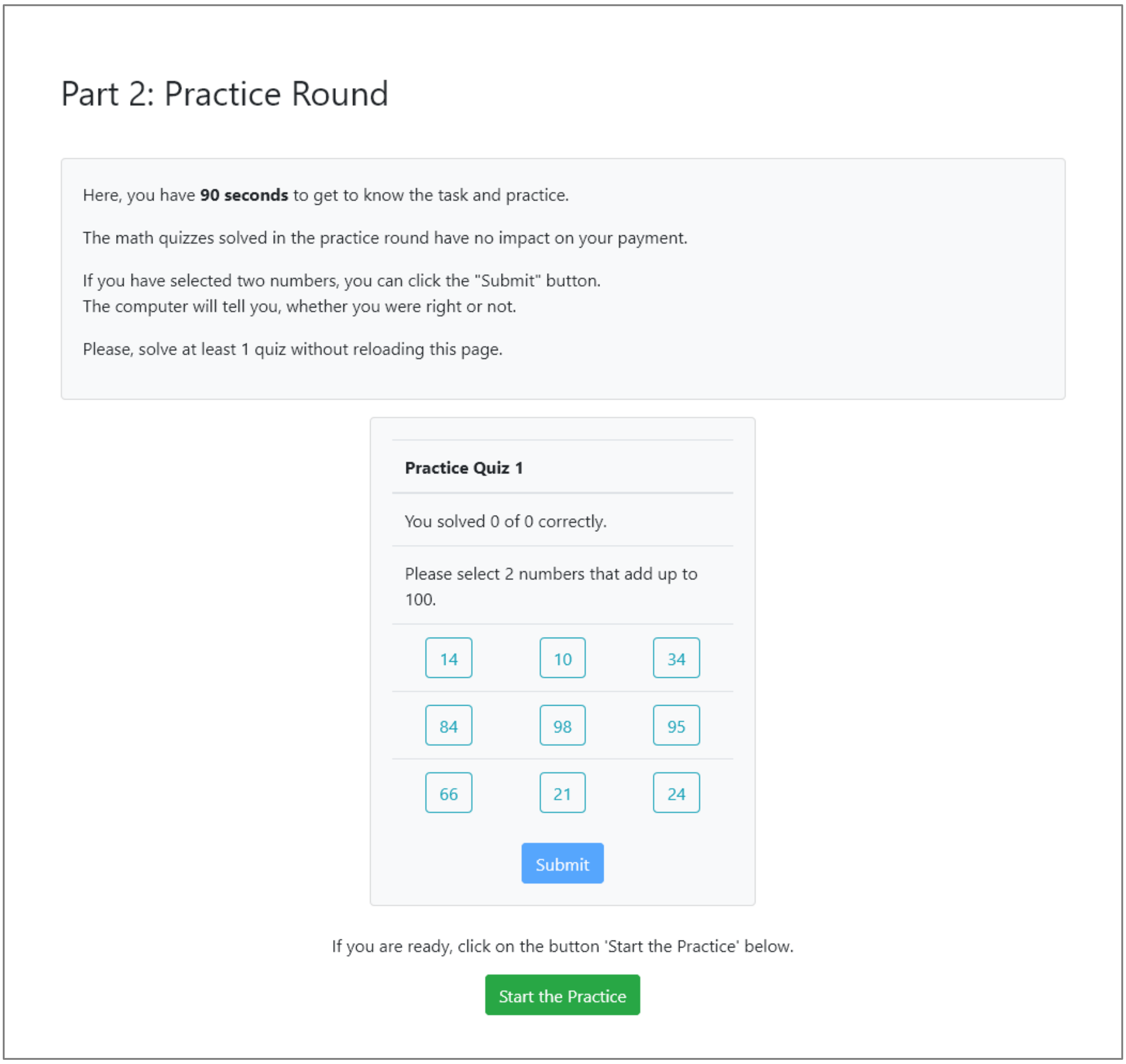
You cannot select more than two numbers. So if you want to change your selection, unselect one of your choices by clicking a second time on your choice and select the one you want.

After submitting an answer, you will get a new quiz. Below the quiz number, you can see how many quizzes you have already solved correctly.

You have a time limit of **120 seconds** to complete as many quizzes as possible in each part.

You can continue to the practice round, where you can practice the exercise.

Continue to Practice Round


Part 2: Practice Round
Here, you have **90 seconds** to get to know the task and practice.
The math quizzes solved in the practice round have no impact on your payment.
If you have selected two numbers, you can click the "Submit" button.
The computer will tell you, whether you were right or not.
Please, solve at least 1 quiz without reloading this page.
Practice Quiz 1
You solved 0 of 0 correctly.
Please select 2 numbers that add up to 100.
14
10
34
84
98
95
66
21
24
Submit
If you are ready, click on the button 'Start the Practice' below.
Start the Practice

Part 2: Practice Round

Time left to practice: **1:21**

Here, you have **90 seconds** to get to know the task and practice.

The math quizzes solved in the practice round have no impact on your payment.

If you have selected two numbers, you can click the "Submit" button.
The computer will tell you, whether you were right or not.

Please, solve at least 1 quiz without reloading this page.

Practice Quiz 2

You solved 1 of 1 correctly.

Please select 2 numbers that add up to 100.

20	13	55
80	92	12
22	54	40

Submit

If you solved 1 quiz, you can continue.

Continue

Part 2: Payment Details

Payment PIECE RATE

- The money you can earn in this part depends only on your own performance.
- You receive **1 point** for each correctly solved quiz.
- You do not loose points if you submit a wrong answer.
- If this part is randomly selected for payment at the end of the study, you get **£0.50** for each collected point.

If you are ready, click on the button 'Start this part' below.
You have 120 seconds and your submissions are relevant for your payment.

Start this part

Part 2: Math Quizzes

Time left to solve quizzes: **1:58**

Quiz 1

You solved 0 of 0 correctly.

Please select 2 numbers that add up to 100.

38	34	15
96	58	31
85	36	63

Submit

Part 2: Bonus Question

Please answer the following bonus question

You now have the chance to earn additional money. If this part is randomly chosen for payment, you will be paid an extra £1.00 on top of your final payoff if your following answer is correct.

Suppose we group all participants of this study into 4 groups by their achieved points and assume we have 100 participants for simplicity. First, we put the 25 participants with the highest amounts of points in Group 1. The next 25 participants with the second highest amounts of points in Group 2 and the following 25 in Group 3. The 25 participants with the fewest amounts of points are in Group 4.

With the points you collected, to which group do you think you belong?

- I would be in Group 1.
- I would be in Group 2.
- I would be in Group 3.
- I would be in Group 4.

Continue to Summary

Part 2: Summary

Result

You solved **1** of **1** quizzes correctly.

Continue to next part

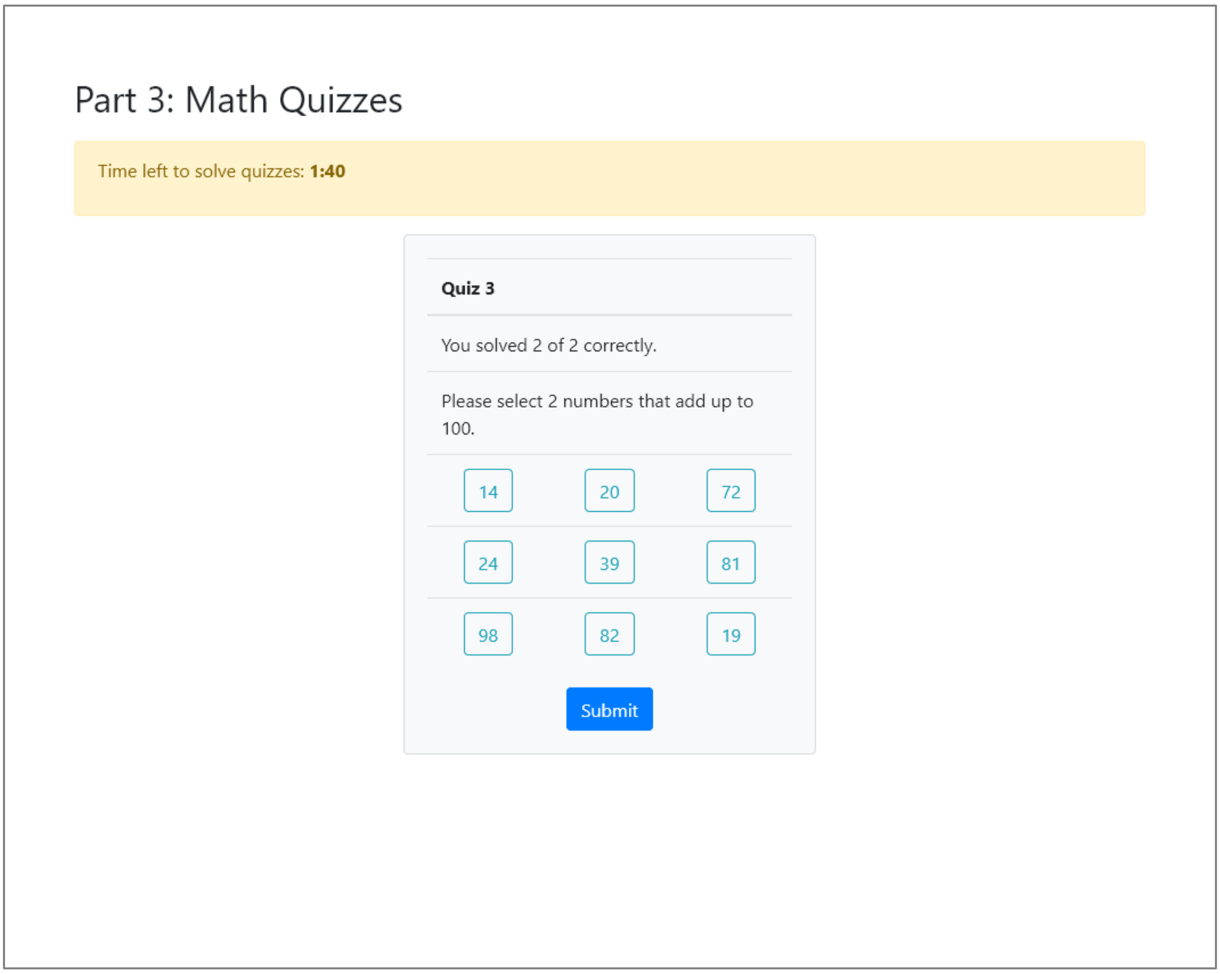
Part 3: Payment Details

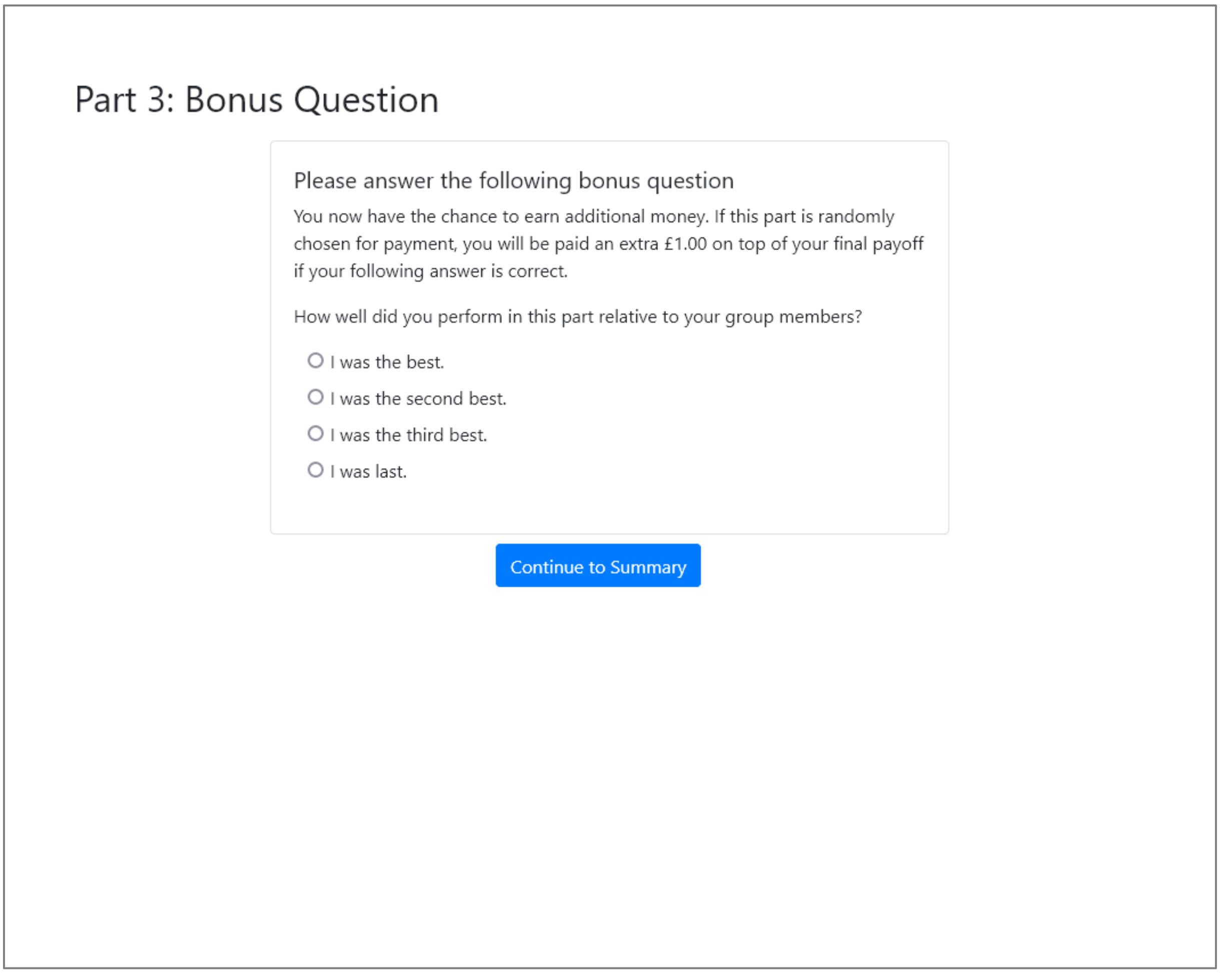
Payment TOURNAMENT

- For this part, you are now randomly matched with **three other participants** of this study. You are **one group**. You will not get to know who the other participants are and the other participants will not get to know who you are.
- The money you can earn in this part depends on your performance and the performance of the other group members. Thus, you are in a **tournament** with the other group members.
- You get again **1 point** for each correctly solved quiz and you do not loose points if you submit a wrong answer.
- If this part is randomly selected for payment at the end of the study, your payment is determined as follows:
  - If you collect *more points* than all other members of your group, you are the winner of this tournament and get **£2.00** for each point.
  - If you collect *less points* than the best player in your group, you get **nothing** in this part.
  - If there are ties, the winner will be randomly determined.

If you are ready, click on the button 'Start this part' below.
You have 120 seconds and your submissions are relevant for your payment.

Start this part


Part 3: Math Quizzes
Time left to solve quizzes: 1:40
Quiz 3
You solved 2 of 2 correctly.
Please select 2 numbers that add up to 100.
14
20
72
24
39
81
98
82
19
Submit


Part 3: Bonus Question
Please answer the following bonus question
You now have the chance to earn additional money. If this part is randomly chosen for payment, you will be paid an extra £1.00 on top of your final payoff if your following answer is correct.
How well did you perform in this part relative to your group members?
I was the best.
I was the second best.
I was the third best.
I was last.
Continue to Summary

Part 3: Summary

Result

You solved **2** of **2** quizzes correctly.

Continue to next part

Part 4: Your Choice

Please choose one of the two options

In this part, you can choose which payment scheme from the previous two parts should apply to your performance.

You can choose **Option A: PIECE RATE** from Part 2 or **Option B: TOURNAMENT** from Part 3.

Below you can see both options, and the respective resulting payment rules.

Option A: PIECE RATE

Payment

- The money you can earn in this part depends only on your own performance.
- You receive **1 point** for each correctly solved quiz.
- You do not loose points if you submit a wrong answer.
- If this part is randomly selected for payment at the end of the study, you get **£0.50** for each collected point.

Option B: TOURNAMENT

If you choose Option B, we compare your points to the points of your group members from the previous part. Hence, we take your points from this part and compare them with the other group members' points in Part 3. This means that your choice in this part has no impact on the payment of the other group members.

Payment

- You get again **1 point** for each correctly solved quiz and you do not loose points if you submit a wrong answer.
- If this part is randomly selected for payment at the end of the study, your payment is determined as follows:
  - If you collect *more points* than all other members of your group in Part 3, you are the winner of this tournament and get **£2.00** for each point.
  - If you collect *less points* than the best player in your group in Part 3, you get **nothing** in this part.
  - If there are ties, the winner will be randomly determined.

Please choose:

Option A or Option B

Choice: Option A

Part 4: Payment Details

You have chosen **Option A**.

Payment

- The money you can earn in this part depends only on your own performance.
- You receive **1 point** for each correctly solved quiz.
- You do not loose points if you submit a wrong answer.
- If this part is randomly selected for payment at the end of the study, you get **£0.50** for each collected point.

If you are ready, click on the button 'Start this part' below.
You have 120 seconds and your submissions are relevant for your payment.

Start this part

Choice: Option B

Part 4: Payment Details

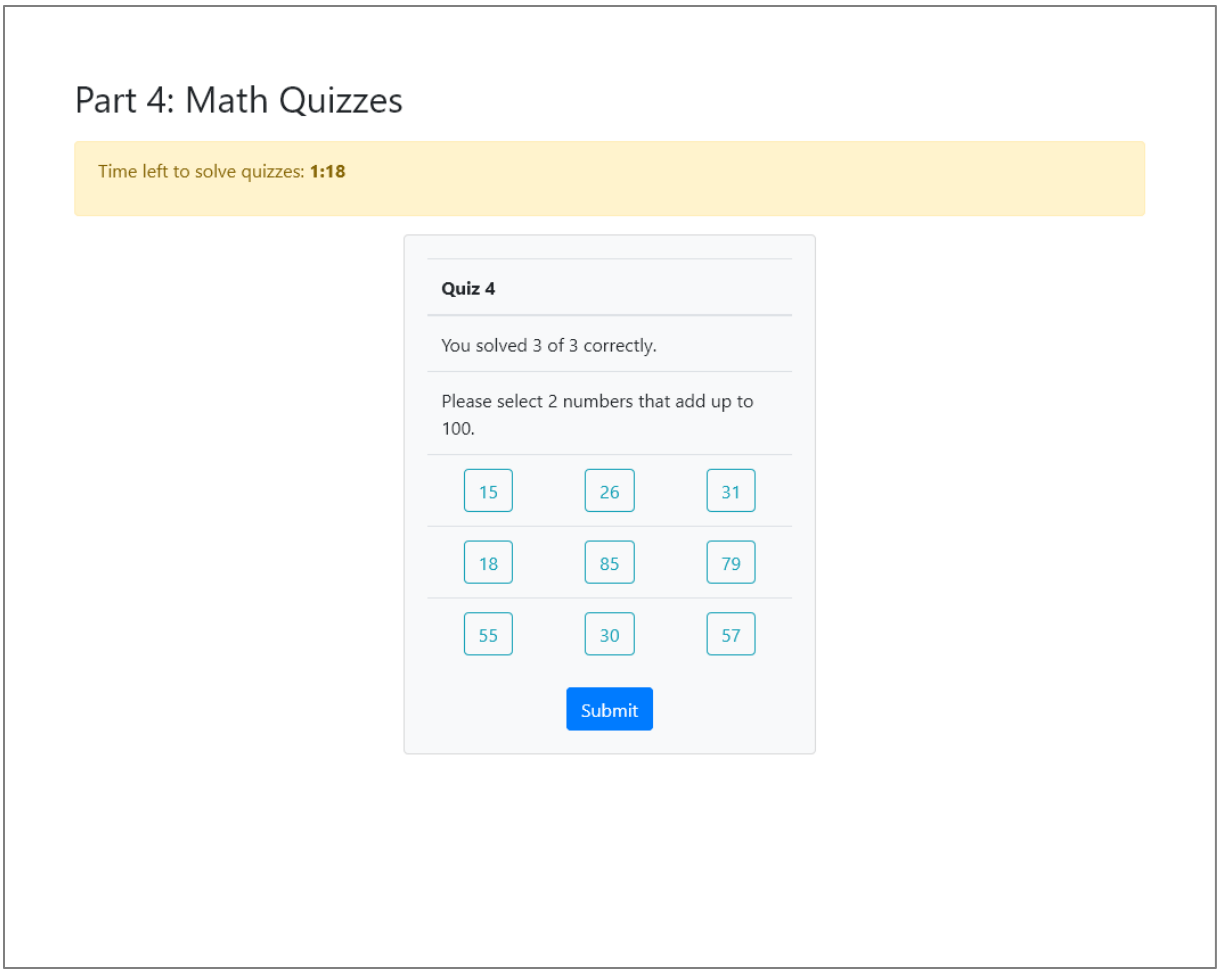
You have chosen **Option B**.

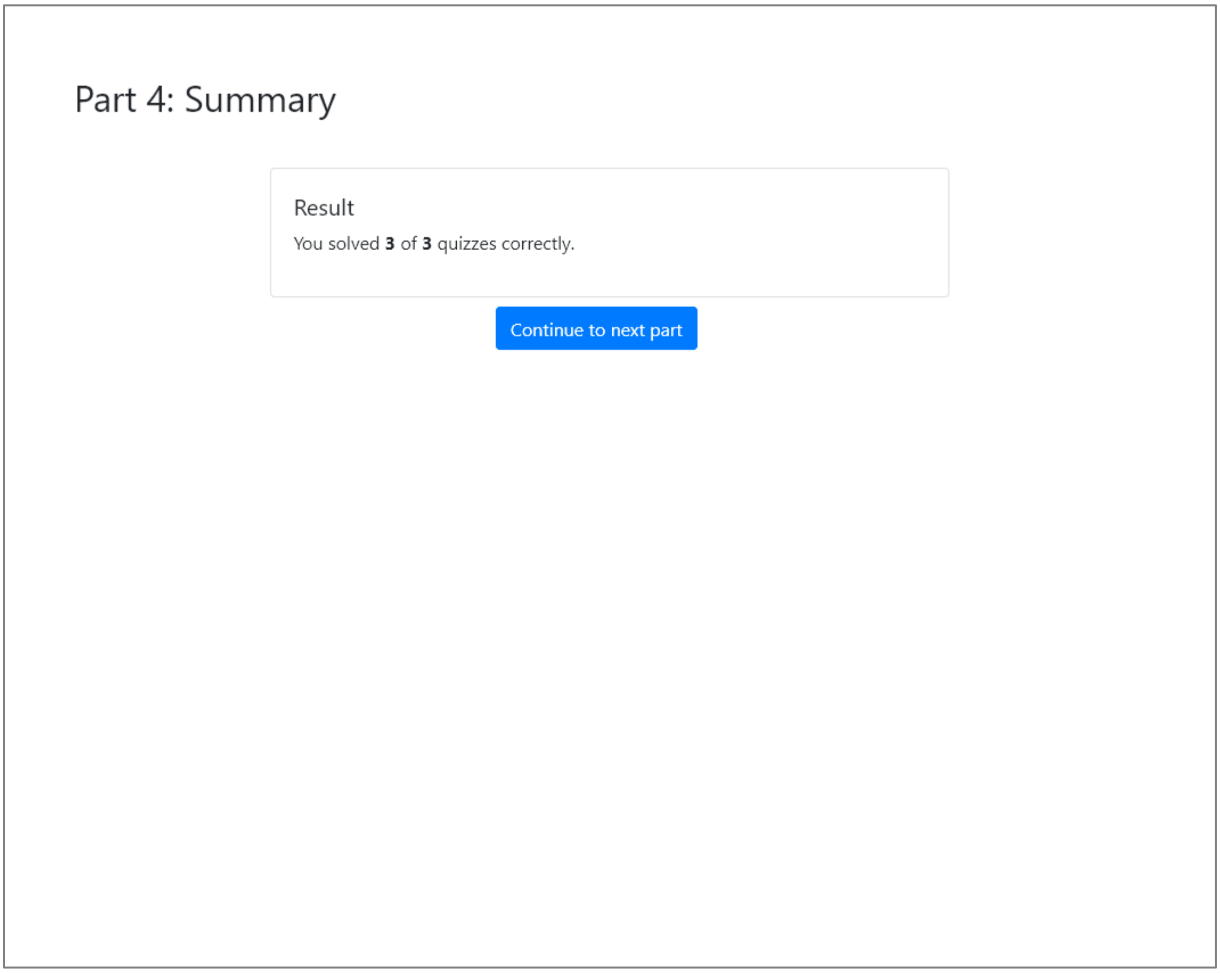
Payment

- You get again **1 point** for each correctly solved quiz and you do not loose points if you submit a wrong answer.
- If this part is randomly selected for payment at the end of the study, your payment is determined as follows:
  - If you collect *more points* than all other members of your group in Part 3, you are the winner of this tournament and get **£2.00** for each point.
  - If you collect *less points* than the best player in your group in Part 3, you get **nothing** in this part.
  - If there are ties, the winner will be randomly determined.

If you are ready, click on the button 'Start this part' below.
You have 120 seconds and your submissions are relevant for your payment.

Start this part


Part 4: Math Quizzes
Time left to solve quizzes: 1:18
Quiz 4
You solved 3 of 3 correctly.
Please select 2 numbers that add up to 100.
15
26
31
18
85
79
55
30
57
Submit


Part 4: Summary
Result
You solved 3 of 3 quizzes correctly.
Continue to next part

Part 5: Instructions

In this part, you get a start capital of **£4.00** and you have to decide how much of it you want to invest into a risky lottery.

The success of this investment is decided by *flipping a virtual coin*. You can get 'heads' or 'tails' with an equal probability. That means you have a 50% chance of success.

If you get **'tails'**, your investment was successful. In this case, the amount you have invested is multiplied by **2.5**. Your earnings are then the leftover of your start capital and the money from the successful investment.

But if you get **'heads'**, your investment was not successful, and you have only the leftover of your start capital.

There are no restrictions: you can put everything of your start capital in the investment, or nothing, or anything in between.

On the next page, you can decide how much to invest and see the potential outcomes of your investment.

Continue

Part 5: Your Decision

Your Decision

How much of your £4.00 do you want to invest?

4 £

If you have set the amount you want to invest, click on the button "Invest" below.

Under "Potential Outcome" to the right, you can see what can happen when you click "Invest".

Potential Outcome

You want to invest: £4.00

You want to keep: £0.00

If the coin shows 'tails', you additionally get £10.00.
So in total you get: £0.00 + £10.00 = **£10.00**

If the coin shows 'heads', you get £0.
So in total you get: £0.00 + £0 = **£0.00**

If you have made your decision, please click on the button 'Invest' below.

Invest

Part 5: Summary

You started with a capital of **£4.00**.

You invested **£4.00**.

We are flipping the virtual coin to see if your investment is a success or not. You will see the result at the end of the experiment.

Continue to next Part

Part 6: Instructions

You will receive an initial endowment of **£5.00** for this part.

On the next page, you can decide how much of this endowment you want to keep for yourself and how much you would like to donate to five different charities.

You can freely set the share you want to keep and the share you want to transfer to one or more respective charities. Any amount between 0 and £5.00 is possible, as long as the sum of all options equals your endowment.

Once we have finished the study, we will send the respective donations of all participants to the charities.

Continue

Part 6: Your Donation

Please indicate the amount you want to donate.

You have received an initial endowment of £5.00.

Remember that the sum must be equal to your endowment.

Amnesty International:

1 £

Doctors without Borders:

1 £

International Red Cross:

1 £

UNICEF:

1 £

WWF:

1 £

Your share:

0.00 £

Submit

Part 6: Summary

You started with an endowment of **£5.00**.

You kept for yourself **£0.00**.

You donated **£5.00**.

Continue to next Part

Questionnaire

You have finished the six parts.

Now, we kindly ask you to answer the following questions.

At the end, you will get your personal unique **8-digit Code** for the completion of the study.

Start Questionnaire

Questionnaire

Below you find several personality traits that may or may not apply to you.

Please indicate for each trait the extent to which it applies to you on a scale from 1 (*Never true*) to 6 (*Always true*).

Trait	*Never true*	*Sometimes true*	*Occasionally true*	*Often true*	*Usually true*	*Always true*
Defends own beliefs	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Tender	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Conscientious	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Independent	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Sympathetic	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Moody	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Assertive	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Sensitive to needs of other	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Reliable	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Strong personality	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Understanding	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Jealous	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Forceful	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Compassionate	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Truthful	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6

Continue

Questionnaire

Below you find several personality traits that may or may not apply to you.

Please indicate for each trait the extent to which it applies to you on a scale from 1 (*Never true*) to 6 (*Always true*).

Trait	*Never true*	*Sometimes true*	*Occasionally true*	*Often true*	*Usually true*	*Always true*
Has leader abilities	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Eager to soothe hurt feelings	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Secretive	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Willing to take risk	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Warm	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Adaptable	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Dominant	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Affectionate	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Conceited	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Willing to take a stand	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Loves children	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Tactful	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Aggressive	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Gentle	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6
Conventional	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6

Continue

Questionnaire

Gender identity is defined as the gender that you identify yourself with. It is not necessarily related to your assigned sex at birth.

For the following statements, please indicate the response that best describes your experience over the past two weeks on a scale from 1 (*Strongly disagree*) to 5 (*Strongly agree*).

My outward appearance represents my gender identity.						
Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*
I experience a sense of unity between my gender identity and my body.						
Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*
My physical appearance adequately expresses my gender identity.						
Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*
I am generally comfortable with how others perceive my gender identity when they look at me.						
Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*
My physical body represents my gender identity.						
Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*
The way my body currently looks does not represent my gender identity.						
Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*

Continue

Questionnaire

Gender identity is defined as the gender that you identify yourself with. It is not necessarily related to your assigned sex at birth.

For the following statements, please indicate the response that best describes your experience over the <u>past two weeks</u> on a scale from 1 (*Strongly disagree*) to 5 (*Strongly agree*).

I am happy with the way my appearance expresses my gender identity.

Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*

I do not feel that my appearance reflects my gender identity.

Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*

I feel that my mind and body are consistent with one another.

Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*

I am not proud of my gender identity.

Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*

I am happy that I have the gender identity that I do.

Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*

I have accepted my gender identity.

Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	*Strongly agree*

Continue

Questionnaire

Please answer the following questions.

Please indicate your age.

What is you body height in cm (for a converter see: https://www.unitconverters.net/length/feet-to-cm.htm)?

Which is the highest level of education you have completed?

What is your personal income per year (after tax) in GBP?

Are you a student?

- Yes
- No

Where are you currently living?

What is your religious affiliation?

What is your biological sex?

What is your gender?

If you selected "Other", please specify:

Who do you feel attracted to: men, women, both or neither?

Are you currently in a relationship and if yes with a man or a woman?

Do you consider yourself transgender?

- Yes
- No

In what kind of accommodation are you currently living?

Apartment House Dormitory Other

Do you currently live with your partner in the same accommodation?

- Yes
- No

Do you currently live in the same city as your parents?

- Yes
- No

Do you currently live with your parents in the same house/apartment?

- Yes
- No

Do you think your parents treated you mostly as a boy or as girl?

Boy Girl They treated me mostly gender neutral.

On which device do you participate in this study?

Tablet Smartphone PC Laptop Other device

Send and Continue

Questionnaire

Please answer the following questions.

What do you think was the purpose of this study? What do you think the study tried to find out?

Did you think any of the previous tasks were related?

○ Yes ○ No

Did anything you do in one task affect what you did in the other tasks?

○ Yes ○ No

Did you ever see or complete a word-search puzzle in another study?

○ Yes ○ No

Do you remember any of the words from the word-search puzzle? If not leave empty.

Did any of the words in the word-search puzzle seem strange or suspicious to you? If not leave empty.

Send and Continue

Questionnaire

Please indicate whether you have taken any of the following actions in order to transition to your gender identity or if it does not apply.

Come out as transgender to family.

○ Yes ○ No ○ Does not apply

Come out as transgender to friends.

○ Yes ○ No ○ Does not apply

Come out as transgender to coworkers or fellow students.

○ Yes ○ No ○ Does not apply

Adopted a name not given at birth that better represents gender identity.

○ Yes ○ No ○ Does not apply

Currently called adopted name by family.

○ Yes ○ No ○ Does not apply

Currently called adopted name by friends.

○ Yes ○ No ○ Does not apply

Currently called adopted name by coworkers or fellow students.

○ Yes ○ No ○ Does not apply

Legally had name change to adopted name.

○ Yes ○ No ○ Does not apply

Wear clothing that matches gender identity in social situations.

○ Yes ○ No ○ Does not apply

Wear clothing that matches gender identity in work/school.

○ Yes ○ No ○ Does not apply

Legally changed sex on birth certificate (if live in state where this is possible).

○ Yes ○ No ○ Does not apply

Driver's license changed to reflect gender identity.

○ Yes ○ No ○ Does not apply

Had surgery to alter genitalia.

○ Yes ○ No ○ Does not apply

Undergoing hormone replacement therapy.

○ Yes ○ No ○ Does not apply

Used or had a nonsurgical cosmetic procedure (e.g., electrolysis) to alter physical appearance in order to make it more congruent with gender identity.

○ Yes ○ No ○ Does not apply

Had non-genital surgery (e.g., breast removal, breast implants, facial femininization surgery, vocal cord surgery) to alter appearance (or presence) in order to make it more congruent with gender identity.

○ Yes ○ No ○ Does not apply

Send and Continue

Selection of your Payment

Your results in each part

Here you can see the outcomes of all six parts.

The computer will randomly select one of these parts on the next screen. Each part has the same probability to be selected.

You will get **£3.75** and the additional money in the column 'payoff' from this part.

	Short summary of the parts		
Part 1	You found:	Payment:	Payoff
	8 words	£5.00 if all found	£5.00
Part 2	You solved:	Payment:	Payoff
	1 quizzes	£0.50 per quiz	£0.50
Part 3	You solved:	Payment:	Payoff
	2 quizzes	£2.00 per quiz only if you are the best	£4.00
Part 4	You solved:	Payment:	Payoff
	3 quizzes	*Option A* £0.50 per quiz	£1.50
Part 5	You invested:	Success?	Payoff
	£4.00	Yes	£10.00
Part 6	You kept:	You donated:	Payoff
	£0.00	£5.00	£0.00

Continue

Thank you very much for participating!

Please write down or take a screenshot of the following information and the 8-digit code below.

The computer selected:

Part 2

Payment	Part payment	**Total payment**
£3.75	£0.50	£4.25

Your personal code:

4588DDFC

Important: Click on the link below to be redirected to Prolific. Clicking is necessary to claim your reward for participating in this study.
https://app.prolific.co/submissions/complete?cc=4588DDFC

Bibliography

Adams, Renée B and Patricia Funk (2012). "Beyond the glass ceiling: Does gender matter?" In: *Management Science* 58.2, pp. 219–235. DOI: 10.1287/mnsc.1110.1452 (cit. on p. 182).

Aggarwal, Sanjeev Kumar, Lalit Mohan Saini, and Ashwani Kumar (2009). "Electricity price forecasting in deregulated markets: A review and evaluation". In: *International Journal of Electrical Power & Energy Systems* 31.1, pp. 13–22. DOI: 10.1016/j.ijepes.2008.09.003 (cit. on p. 22).

Ahmed, Ali M and Osvaldo Salas (2011). "Implicit influences of Christian religious representations on dictator and prisoner's dilemma game decisions". In: *The Journal of Socio-Economics* 40.3, pp. 242–246. DOI: 10.1016/j.socec.2010.12.013 (cit. on p. 184).

Alempaki, Despoina, Chris Starmer, and Fabio Tufano (2019). "On the priming of risk preferences: The role of fear and general affect". In: *Journal of Economic Psychology* 75, p. 102137. DOI: 10.1016/j.joep.2018.12.011 (cit. on p. 183).

Alhelou, Hassan, Mohamad-Esmail Hamedani-Golshan, Reza Zamani, Ehsan Heydarian-Forushani, and Pierluigi Siano (2018). "Challenges and Opportunities of Load Frequency Control in Conventional, Modern and Future Smart Power Systems: A Comprehensive Review". In: *Energies* 11.10, p. 2497. DOI: 10.3390/en11102497 (cit. on p. 33).

Allen, Eric J., Patricia M. Dechow, Devin G. Pope, and George Wu (2017). "Reference-Dependent Preferences: Evidence from Marathon Runners". In: *Management Science* 63.6, pp. 1657–1672. DOI: 10.1287/mnsc.2015.2417 (cit. on p. 55).

Almås, Ingvild, Alexander W Cappelen, Kjell G Salvanes, Erik Ø Sørensen, and Bertil Tungodden (2016a). "What explains the gender gap in college track dropout? Experimental and administrative evidence". In: *American Economic Review* 106.5, pp. 296–302. DOI: 10.1257/aer.p20161075 (cit. on p. 77).

Almås, Ingvild, Alexander W. Cappelen, Kjell G. Salvanes, Erik Ø. Sørensen, and Bertil Tungodden (2016b). "Willingness to compete: Family matters". In: *Management Science* 62.8, pp. 2149–2162. DOI: 10.1287/mnsc.2015.2244 (cit. on pp. 77, 82, 84, 85, 94).

Almenberg, Johan and Anna Dreber (2015). "Gender, stock market participation and financial literacy". In: *Economics Letters* 137, pp. 140–142. DOI: 10.1016/j.econlet.2015.10.009 (cit. on p. 182).

American Psychological Association (2015). “Guidelines for psychological practice with transgender and gender nonconforming people”. In: *American Psychologist* 70.9, pp. 832–864. DOI: 10.1037/a0039906 (cit. on p. 86).

Andersen, Steffen, Seda Ertac, Uri Gneezy, John A. List, and Sandra Maximiano (2013). “Gender, competitiveness, and socialization at a young age: Evidence from a matrilineal and a patriarchal society”. In: *Review of Economics and Statistics* 95.4, pp. 1438–1443. DOI: 10.1162/REST_a_00312 (cit. on pp. 77, 80, 84).

Anderson, Anders, Anna Dreber, and Roine Vestman (2015). “Risk taking, behavioral biases and genes: Results from 149 active investors”. In: *Journal of Behavioral and Experimental Finance* 6, pp. 93–100. DOI: 10.1016/j.jbef.2015.04.002 (cit. on pp. 85, 182).

Anderson, Lisa R and Jennifer M Mellor (2008). “Predicting health behaviors with an experimental measure of risk preference”. In: *Journal of Health Economics* 27.5, pp. 1260–1274. DOI: 10.1016/j.jhealeco.2008.05.011 (cit. on p. 182).

Apicella, Coren L., Justin M. Carré, and Anna Dreber (2015). “Testosterone and Economic Risk Taking: A Review”. In: *Adaptive Human Behavior and Physiology* 1.3, pp. 358–385. DOI: 10.1007/s40750-014-0020-2 (cit. on p. 84).

Apicella, Coren L., Elif E. Demiral, and Johanna Mollerstrom (2017). “No Gender Difference in Willingness to Compete When Competing against Self”. In: *American Economic Review* 107.5, pp. 136–140. DOI: 10.1257/aer.p20171019 (cit. on p. 84).

Atkinson, Stanley M, Samantha Boyce Baird, and Melissa B Frye (2003). “Do female mutual fund managers manage differently?” In: *Journal of Financial Research* 26.1, pp. 1–18. DOI: 10.1111/1475-6803.00041 (cit. on p. 182).

Aue, Alexander, Diogo Dubart Norinho, and Siegfried Hörmann (2015). “On the Prediction of Stationary Functional Time Series”. In: *Journal of the American Statistical Association* 110.509, pp. 378–392. DOI: 10.1080/01621459.2014.909317 (cit. on pp. 29–31).

Backus, Matthew, Thomas Blake, Brad Larsen, and Steven Tadelis (2020). “Sequential Bargaining in the Field: Evidence from Millions of Online Bargaining Interactions*”. In: *The Quarterly Journal of Economics* 135.3, pp. 1319–1361. DOI: 10.1093/qje/qjaa003 (cit. on pp. 2, 53, 55–58, 113, 115, 116).

Backus, Matthew, Thomas Blake, and Steven Tadelis (2019). “On the empirical content of cheap-talk signaling: An application to bargaining”. In: *Journal of Political Economy* 127.4, pp. 1599–1628. DOI: 10.1086/701699 (cit. on pp. 55, 57).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *ICLR 2015*. Ed. by Yoshua Bengio and Yann LeCun. arXiv: 1409.0473 (cit. on pp. 24, 25, 41, 103).

Balafoutas, Loukas, Helena Fornwagner, and Matthias Sutter (2018). “Closing the gender gap in competitiveness through priming”. In: *Nature Communications* 9.1, pp. 1–6. DOI: 10.1038/s41467-018-06896-6 (cit. on pp. 82, 85).

Balafoutas, Loukas, Rudolf Kerschbamer, and Matthias Sutter (2012). "Distributional preferences and competitive behavior". In: *Journal of Economic Behavior & Organization* 83.1, pp. 125–135. DOI: 10.1016/j.jebo.2011.06.018 (cit. on p. 84).

Balafoutas, Loukas and Matthias Sutter (2012). "Affirmative action policies promote women and do not harm efficiency in the laboratory". In: *Science* 335.6068, pp. 579–582. DOI: 10.1126/science.1211180 (cit. on pp. 82, 85).

Balafoutas, Loukas and Matthias Sutter (2019). "How uncertainty and ambiguity in tournaments affect gender differences in competitive behavior". In: *European Economic Review* 118, pp. 1–13. DOI: 10.1016/j.euroecorev.2019.05.005 (cit. on p. 77).

Baldiga, Nancy R and Katherine B Coffman (2018). "Laboratory evidence on the effects of sponsorship on the competitive preferences of men and women". In: *Management Science* 64.2, pp. 888–901. DOI: 10.1287/mnsc.2016.2606. (cit. on p. 82).

Banerjee, Ritwik, Nabanita Datta Gupta, and Marie Claire Villeval (2018). "The spillover effects of affirmative action on competitiveness and unethical behavior". In: *European Economic Review* 101, pp. 567–604. DOI: 10.1016/j.euroecorev.2017.10.022 (cit. on p. 84).

Bardsley, Nicholas, Judith Mehta, Chris Starmer, and Robert Sugden (2010). "Explaining focal points: Cognitive hierarchy theory versus team reasoning". In: *The Economic Journal* 120.543, pp. 40–79. DOI: 10.1111/j.1468-0297.2009.02304.x (cit. on p. 54).

Bargh, John A, Peter M Gollwitzer, Annette Lee-Chai, Kimberly Barndollar, and Roman Trötschel (2001). "The automated will: Nonconscious activation and pursuit of behavioral goals." In: *Journal of Personality and Social Psychology* 81.6, pp. 1014–1027. DOI: 10.1037/0022-3514.81.6.1014 (cit. on p. 87).

Barsky, Robert B, F Thomas Juster, Miles S Kimball, and Matthew D Shapiro (1997). "Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study". In: *The Quarterly Journal of Economics* 112.2, pp. 537–579. DOI: 10.1162/003355397555280 (cit. on p. 182).

Bateson, Melissa, Daniel Nettle, and Gilbert Roberts (2006). "Cues of being watched enhance cooperation in a real-world setting". In: *Biology Letters* 2.3, pp. 412–414. DOI: 10.1098/rsbl.2006.0509 (cit. on p. 184).

Beblo, Miriam and Eva Markowsky (2022). "When do we observe a gender gap in competition entry? A meta-analysis of the experimental literature". In: *Journal of Economic Behavior and Organization* Conditionally accepted (cit. on p. 85).

Benjamin, Daniel J, James J Choi, and Geoffrey Fisher (2016). "Religious identity and economic behavior". In: *Review of Economics and Statistics* 98.4, pp. 617–637. DOI: 10.1162/REST_a_00586 (cit. on p. 184).

Benjamin, Daniel J, James J Choi, and A Joshua Strickland (2010). "Social identity and preferences". In: *American Economic Review* 100.4, pp. 1913–28. DOI: 10.1257/aer.100.4.1913 (cit. on p. 183).

Benz, Matthias and Stephan Meier (2008). “Do people behave in experiments as in the field?—evidence from donations”. In: *Experimental Economics* 11.3, pp. 268–281. DOI: 10.1007/s10683-007-9192-y (cit. on p. 183).

Berge, Lars Ivar Oppedal, Kjetil Bjorvatn, Armando Jose Garcia Pires, and Bertil Tungodden (2015). “Competitive in the lab, successful in the field?” In: *Journal of Economic Behavior & Organization* 118, pp. 303–317. DOI: 10.1016/j.jebo.2014.11.014 (cit. on p. 77).

Berlin, Noémi and Marie-Pierre Dargnies (2016). “Gender differences in reactions to feedback and willingness to compete”. In: *Journal of Economic Behavior & Organization* 130, pp. 320–336. DOI: 10.1016/j.jebo.2016.08.002 (cit. on p. 84).

Besse, Philippe C., Herve Cardot, and David B. Stephenson (2000). “Autoregressive Forecasting of Some Functional Climatic Variations”. In: *Scandinavian Journal of Statistics* 27.4, pp. 673–687. DOI: 10.1111/1467-9469.00215 (cit. on p. 30).

Bilén, David, Anna Dreber, and Magnus Johannesson (2021). “Are women more generous than men? A meta-analysis”. In: *Journal of the Economic Science Association* 7.1, pp. 1–18. DOI: 10.1007/s40881-021-00105-9 (cit. on pp. 85, 183).

Black, Fischer and Myron Scholes (1973). “The Pricing of Options and Corporate Liabilities”. In: *Journal of Political Economy* 81.3, pp. 637–654. DOI: 10.1086/260062 (cit. on p. 14).

Boksem, Maarten AS, Pranjal H Mehta, Bram Van den Bergh, Veerle van Son, Stefan T Trautmann, Karin Roelofs, Ale Smidts, and Alan G Sanfey (2013). “Testosterone inhibits trust but promotes reciprocity”. In: *Psychological Science* 24.11, pp. 2306–2314. DOI: 10.1177/0956797613495063 (cit. on pp. 85, 182).

Bollerslev, Tim (1986). “Generalized autoregressive conditional heteroskedasticity”. In: *Journal of Econometrics* 31.3, pp. 307–327. DOI: 10.1016/0304-4076(86)90063-1 (cit. on pp. 28, 31).

Bollerslev, Tim (1987). “A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return”. In: *The Review of Economics and Statistics* 69.3, p. 542. JSTOR: 1925546. DOI: 10.2307/1925546 (cit. on p. 29).

Bonin, Holger, Thomas Dohmen, Armin Falk, David Huffman, and Uwe Sunde (2007). “Cross-sectional earnings risk and occupational sorting: The role of risk attitudes”. In: *Labour Economics* 14.6, pp. 926–937. DOI: 10.1016/j.labeco.2007.06.007 (cit. on p. 182).

Bönte, Werner, Sandro Lombardo, and Diemo Urbig (2017). “Economics meets psychology: Experimental and self-reported measures of individual competitiveness”. In: *Personality and Individual Differences* 116, pp. 179–185. DOI: 10.1016/j.paid.2017.04.036 (cit. on p. 84).

Bönte, Werner, Vivien Procher, and Diemo Urbig (2018). “Gender differences in selection into self-competition”. In: *Applied Economics Letters* 25.8, pp. 539–543. DOI: 10.1080/13504851.2017.1343441 (cit. on p. 84).

Boomsma, Krogh Trine, Nina Juul, and Stein-Erik Fleten (2014). "Bidding in sequential electricity markets: The Nordic case". In: *European Journal of Operational Research* 238.3, pp. 797–809. DOI: 10.1016/j.ejor.2014.04.027 (cit. on p. 23).

Booth, Alison, Elliott Fan, Xin Meng, and Dandan Zhang (2019). "Gender Differences in Willingness to Compete: The Role of Culture and Institutions". In: *The Economic Journal* 129.618, pp. 734–764. DOI: 10.1111/ecoj.12583 (cit. on p. 84).

Booth, Alison and Patrick Nolen (2012). "Choosing to compete: How different are girls and boys?" In: *Journal of Economic Behavior & Organization* 81.2, pp. 542–555. DOI: 10.1016/j.jebo.2011.07.018 (cit. on p. 82).

Borne, Olivier, Klaas Korte, Yannick Perez, Marc Petit, and Alexandra Purkus (2018). "Barriers to entry in frequency-regulation services markets: Review of the status quo and options for improvements". In: *Renewable and Sustainable Energy Reviews* 81, pp. 605–614. DOI: 10.1016/j.rser.2017.08.052 (cit. on pp. 23, 24).

Boschini, Anne, Anna Dreber, Emma von Essen, Astri Muren, and Eva Ranehill (2018). "Gender and altruism in a random sample". In: *Journal of Behavioral and Experimental Economics* 77, pp. 72–77. DOI: 10.1016/j.socec.2018.09.005 (cit. on p. 184).

Bosq, Denis (2000). *Linear processes in function spaces: theory and applications*. Lecture Notes in Statistics 149. New York: Springer. 283 pp. (cit. on pp. 29, 31).

Brandts, Jordi, Valeska Groenert, and Christina Rott (2015). "The Impact of Advice on Women's and Men's Selection into Competition". In: *Management Science* 61.5, pp. 1018–1035. DOI: 10.1287/mnsc.2013.1877 (cit. on pp. 82, 84).

Bucciol, Alessandro and Raffaele Miniaci (2011). "Household portfolios and implicit risk preference". In: *Review of Economics and Statistics* 93.4, pp. 1235–1250. DOI: 10.1162/REST_a_00138 (cit. on p. 182).

Buser, Thomas, Anna Dreber, and Johanna Mollerstrom (2017a). "The impact of stress on tournament entry". In: *Experimental Economics* 20.2, pp. 506–530. DOI: 10.1007/s10683-016-9496-x (cit. on p. 84).

Buser, Thomas, Lydia Geijtenbeek, and Erik Plug (2018a). "Sexual orientation, competitiveness and income". In: *Journal of Economic Behavior & Organization* 151, pp. 191–198. DOI: 10.1016/j.jebo.2018.03.017 (cit. on pp. 77, 83–85, 183).

Buser, Thomas, Leonie Gerhards, and Joël van der Weele (2018b). "Responsiveness to feedback as a personal trait". In: *Journal of Risk and Uncertainty*, pp. 165–192. DOI: 10.1007/s11166-018-9277-3 (cit. on p. 84).

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek (2014). "Gender, Competitiveness, and Career Choices". In: *The Quarterly Journal of Economics* 129.3, pp. 1409–1447. DOI: 10.1093/qje/qju009 (cit. on pp. 77, 84).

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek (2021). *Can competitiveness predict education and labor market outcomes? Evidence from incentivized choice and survey measures*. Working paper 28916. National Bureau of Economic Research. DOI: 10.3386/w28916 (cit. on pp. 77, 87).

Buser, Thomas, Noemi Peter, and Stefan C. Wolter (2017b). “Gender, competitiveness, and study choices in high school: Evidence from Switzerland”. In: *American Economic Review* 107.5, pp. 125–130. DOI: 10.1257/aer.p20171017 (cit. on p. 84).

Buskens, Vincent, Werner Raub, Nynke Van Miltenburg, Estrella R Montoya, and Jack Van Honk (2016). “Testosterone administration moderates effect of social environment on trust in women depending on second-to-fourth digit ratio”. In: *Scientific Reports* 6.1, pp. 1–8. DOI: 10.1038/srep27655 (cit. on pp. 182, 184).

Busse, Meghan R, Nicola Lacetera, Devin G Pope, Jorge Silva-Risso, and Justin R Sydnor (2013). “Estimating the Effect of Salience in Wholesale and Retail Car Markets”. In: *American Economic Review* 103.3, pp. 575–579. DOI: 10.1257/aer.103.3.575 (cit. on pp. 53, 55).

Cadsby, C. Bram, Maroš Servátka, and Fei Song (2013). “How competitive are female professionals? A tale of identity conflict”. In: *Journal of Economic Behavior & Organization* 92, pp. 284–303. DOI: 10.1016/j.jebo.2013.05.009 (cit. on pp. 82, 84).

Callen, Michael, Mohammad Isaqzadeh, James D Long, and Charles Sprenger (2014). “Violence and risk preference: Experimental evidence from Afghanistan”. In: *American Economic Review* 104.1, pp. 123–48. DOI: 10.1257/aer.104.1.123 (cit. on p. 183).

Campos, Fco. Alberto, Antonio Munoz San Roque, Eugenio F. Sanchez-Ubeda, and Jose Portela Gonzalez (2016). “Strategic Bidding in Secondary Reserve Markets”. In: *IEEE Transactions on Power Systems* 31.4, pp. 2847–2856. DOI: 10.1109/TPWRS.2015.2453477 (cit. on p. 23).

Card, David, Ana Rute Cardoso, and Patrick Kline (2016). “Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women”. In: *The Quarterly Journal of Economics* 131.2, pp. 633–686. DOI: 10.1093/qje/qjv038 (cit. on p. 77).

Cárdenas, Juan Camilo, Anna Dreber, Emma von Essen, and Eva Ranehill (2015). “Cooperativeness and competitiveness in children”. In: *Journal of Behavioral and Experimental Economics* 59, pp. 32–41. DOI: 10.1016/j.socec.2015.09.003 (cit. on pp. 82, 85, 182).

Cárdenas, Juan-Camilo, Anna Dreber, Emma von Essen, and Eva Ranehill (2012). “Gender differences in competitiveness and risk taking: Comparing children in Colombia and Sweden”. In: *Journal of Economic Behavior & Organization* 83.1, pp. 11–23. DOI: 10.1016/j.jebo.2011.06.008 (cit. on p. 84).

Carpenter, Jeffrey, Cristina Connolly, and Caitlin Knowles Myers (2008). “Altruistic behavior in a representative dictator experiment”. In: *Experimental Economics* 11.3, pp. 282–298. DOI: 10.1007/s10683-007-9193-x (cit. on p. 183).

Carpenter, Jeffrey, Rachel Frank, and Emiliano Huet-Vaughn (2018). “Gender differences in interpersonal and intrapersonal competitive behavior”. In: *Journal of Behavioral and Experimental Economics* 77, pp. 170–176. DOI: 10.1016/j.socec.2018.10.003 (cit. on p. 84).

Cason, Timothy N., William A. Masters, and Roman M. Sheremeta (2010). “Entry into winner-take-all and proportional-prize contests: An experimental study”. In: *Journal of Public Economics* 94.9-10, pp. 604–611. DOI: 10.1016/j.jpubeco.2010.05.006 (cit. on p. 84).

Cassar, Alessandra and Mary L Rigdon (2021). “Prosocial option increases women’s entry into competition”. In: *Proceedings of the National Academy of Sciences* 118.45. DOI: 10.1073/pnas.2111943118 (cit. on p. 85).

Cassar, Alessandra, Feven Wordofa, and Y. Jane Zhang (2016). “Competing for the benefit of offspring eliminates the gender gap in competitiveness”. In: *Proceedings of the National Academy of Sciences* 113.19, pp. 5201–5205. DOI: 10.1073/pnas.1520235113 (cit. on pp. 82, 84, 85).

Cesarini, David, Christopher T Dawes, Magnus Johannesson, Paul Lichtenstein, and Björn Wallace (2009). “Genetic variation in preferences for giving and risk taking”. In: *The Quarterly Journal of Economics* 124.2, pp. 809–842. DOI: 10.1162/qjec.2009.124.2.809 (cit. on p. 182).

Cesarini, David, Magnus Johannesson, Paul Lichtenstein, Örjan Sandewall, and Björn Wallace (2010). “Genetic variation in financial decision-making”. In: *The Journal of Finance* 65.5, pp. 1725–1754. DOI: 10.1111/j.1540-6261.2010.01592.x (cit. on p. 182).

Cesarini, David, Magnus Johannesson, Patrik KE Magnusson, and Björn Wallace (2012). “The behavioral genetics of behavioral anomalies”. In: *Management Science* 58.1, pp. 21–34. DOI: 10.1287/mnsc.1110.1329 (cit. on pp. 85, 182).

Chan, S.C., K.M. Tsui, H.C. Wu, Yunhe Hou, Yik-Chung Wu, and Felix Wu (2012). “Load/Price Forecasting and Managing Demand Response for Smart Grids: Methodologies and Challenges”. In: *IEEE Signal Processing Magazine* 29.5, pp. 68–85. DOI: 10.1109/MSP.2012.2186531 (cit. on p. 22).

Charness, Gary and Uri Gneezy (2012). “Strong evidence for gender differences in risk taking”. In: *Journal of Economic Behavior & Organization* 83.1, pp. 50–58. DOI: 10.1016/j.jebo.2011.06.007 (cit. on p. 182).

Charness, Gary and Marie-Claire Villeval (2009). “Cooperation and competition in intergenerational experiments in the field and the laboratory”. In: *American Economic Review* 99.3, pp. 956–78. DOI: 10.1257/aer.99.3.956 (cit. on p. 94).

Chartrand, Tanya L and John A Bargh (1996). “Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions.” In: *Journal of Personality and Social Psychology* 71.3, p. 464. DOI: 10.1037/0022-3514.71.3.464 (cit. on p. 88).

Chaudhari, Sneha, Varun Mithal, Gungor Polatkan, and Rohan Ramanath (2020). “An Attentive Survey of Attention Models”. In: *Transactions on Intelligent Systems and Technology*. arXiv: 1904.02874 (cit. on p. 27).

Chen, Daniel L., Martin Schonger, and Chris Wickens (2016). "oTree—An open-source platform for laboratory, online, and field experiments". In: *Journal of Behavioral and Experimental Finance* 9, pp. 88–97. DOI: 10.1016/j.jbef.2015.12.001 (cit. on pp. 65, 87).

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014a). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha: Association for Computational Linguistics, pp. 1724–1734. DOI: 10.3115/v1/D14-1179 (cit. on p. 20).

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio (2014b). "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches". In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha: Association for Computational Linguistics, pp. 103–111. arXiv: 1409.1259 (cit. on pp. 25, 27).

Cohn, Alain, Jan Engelmann, Ernst Fehr, and Michel André Maréchal (2015). "Evidence for countercyclical risk aversion: An experiment with financial professionals". In: *American Economic Review* 105.2, pp. 860–85. DOI: 10.1257/aer.20131314 (cit. on p. 183).

Cohn, Alain, Ernst Fehr, and Michel André Maréchal (2017). "Do professional norms in the banking industry favor risk-taking?" In: *The Review of Financial Studies* 30.11, pp. 3801–3823. DOI: 10.1093/rfs/hhx003 (cit. on p. 183).

Conejo, A.J., M.A. Plazas, R. Espinola, and A.B. Molina (2005). "Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA Models". In: *IEEE Transactions on Power Systems* 20.2, pp. 1035–1042. DOI: 10.1109/TPWRS.2005.846054 (cit. on p. 28).

Converse, Benjamin A. and Patrick J. Dennis (2018). "The role of "Prominent Numbers" in open numerical judgment: Strained decision makers choose from a limited set of accessible numbers". In: *Organizational Behavior and Human Decision Processes* 147, pp. 94–107. DOI: 10.1016/j.obhdp.2018.05.007 (cit. on pp. 53, 55).

Craven, P and G Wahba (1978). "Smoothing noisy data with spline functions". In: *Numerische Mathematik* 31, pp. 377–403 (cit. on p. 10).

Crawford, Vincent P, Uri Gneezy, and Yuval Rottenstreich (2008). "The Power of Focal Points Is Limited: Even Minute Payoff Asymmetry May Yield Large Coordination Failures". In: *American Economic Review* 98.4, pp. 1443–58. DOI: 10.1257/aer.98.4.1443 (cit. on p. 54).

Crespo Cuaresma, Jesús, Jaroslava Hlouskova, Stephan Kossmeier, and Michael Obersteiner (2004). "Forecasting electricity spot-prices using linear univariate time-series models". In: *Applied Energy* 77.1, pp. 87–106. DOI: 10.1016/S0306-2619(03)00096-5 (cit. on p. 28).

Croson, Rachel and Uri Gneezy (2009). "Gender differences in preferences". In: *Journal of Economic Literature* 47.2, pp. 448–474. DOI: 10.1257/jel.47.2.448 (cit. on pp. 70, 85, 182).

Cuijpers, Pim (2016). *Meta-analysis in mental health: A practical guide* (cit. on p. 78).

Czibor, Eszter, Jörg Claussen, and Mirjam Van Praag (2019). "Women in a men's world: Risk taking in an online card game community". In: *Journal of Economic Behavior & Organization* 158, pp. 62–89. DOI: 10.1016/j.jebo.2018.11.011 (cit. on p. 182).

Dargnies, Marie-Pierre (2012). "Men Too Sometimes Shy Away from Competition: The Case of Team Competition". In: *Management Science* 58.11, pp. 1982–2000. DOI: 10.1287/mnsc.1120.1542 (cit. on p. 84).

Dariel, Aurelie, Curtis Kephart, Nikos Nikiforakis, and Christina Zenker (2017). "Emirati women do not shy away from competition: evidence from a patriarchal society in transition". In: *Journal of the Economic Science Association* 3.2, pp. 121–136. DOI: 10.1007/s40881-017-0045-y (cit. on pp. 78, 84).

Datta Gupta, Nabanita, Anders Poulsen, and Marie Claire Villeval (2013). "Gender Matching and Competitiveness: Experimental Evidence". In: *Economic Inquiry* 51.1, pp. 816–835. DOI: 10.1111/j.1465-7295.2011.00378.x (cit. on pp. 77, 84).

De Boor, Carl (2001). *A practical guide to splines*. Revised Edition. Applied Mathematical Sciences 27. New York: Springer. 346 pp. (cit. on pp. 6, 10).

Didericksen, Devin, Piotr Kokoszka, and Xi Zhang (2012). "Empirical properties of forecasts with the functional autoregressive model". In: *Computational Statistics* 27.2, pp. 285–298. DOI: 10.1007/s00180-011-0256-2 (cit. on p. 29).

Dreber, Anna and Magnus Johannesson (2018). *Sex hormones and economic decision making in the lab: A review of the causal evidence*. Routledge, pp. 391–402 (cit. on p. 83).

Dreber, Anna, Emma von Essen, and Eva Ranehill (2014). "Gender and competition in adolescence: task matters". In: *Experimental Economics* 17.1, pp. 154–172. DOI: 10.1007/s10683-013-9361-0 (cit. on p. 84).

Drupp, Moritz A, Menusch Khadjavi, Marie-Catherine Riekhof, and Rudi Voss (2020). "Professional identity and the gender gap in risk-taking. Evidence from field experiments with scientists". In: *Journal of Economic Behavior & Organization* 170, pp. 418–432. DOI: 10.1016/j.jebo.2019.12.020 (cit. on p. 182).

Eckel, Catherine C and Philip J Grossman (1996). "Altruism in anonymous dictator games". In: *Games and Economic Behavior* 16.2, pp. 181–191. DOI: 10.1006/game.1996.0081 (cit. on p. 183).

Eilers, Paul H C, Brian D Marx, and Maria Durban (2015). "Twenty years of P-splines". In: *Statistics and Operations Research Transactions* 39.2, pp. 149–186 (cit. on pp. 11, 39).

Eilers, Paul H. C. and Brian D. Marx (1996). “Flexible smoothing with B -splines and penalties”. In: *Statistical Science* 11.2, pp. 89–121. DOI: 10.1214/ss/1038425655 (cit. on pp. 10, 39).

Eilers, Paul H. C. and Brian D. Marx (2010). “Splines, knots, and penalties”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.6, pp. 637–653. DOI: 10.1002/wics.125 (cit. on p. 11).

Englmaier, Florian, Arno Schmöller, and Till Stowasser (2018). “Price Discontinuities in an Online Market for Used Cars”. In: *Management Science* 64.6, pp. 2754–2766. DOI: 10.1287/mnsc.2016.2714 (cit. on pp. 53, 55).

Erb, Hans-Peter, Antoine Bioy, and Denis J Hilton (2002). “Choice preferences without inferences: Subconscious priming of risk attitudes”. In: *Journal of Behavioral Decision Making* 15.3, pp. 251–262. DOI: 10.1002/bdm.416 (cit. on p. 183).

Fay, Michael P. and Michael A. Proschan (2010). “Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules”. In: *Statistics Surveys* 4 (none). DOI: 10.1214/09-SS051 (cit. on pp. 12, 16).

Fernández, Carmen and Mark F. J. Steel (1998). “On Bayesian modeling of fat tails and skewness”. In: *Journal of the American Statistical Association* 93.441, pp. 359–371. DOI: 10.1080/01621459.1998.10474117 (cit. on p. 29).

Flory, Jeffrey A, Uri Gneezy, Kenneth L Leonard, and John A List (2018). “Gender, age, and competition: A disappearing gap?” In: *Journal of Economic Behavior & Organization* 150, pp. 256–276. DOI: 10.1016/j.jebo.2018.03.027 (cit. on p. 94).

Forbes, C. S., ed. (2011). *Statistical distributions*. 4th ed. Hoboken, N.J: Wiley. 212 pp. (cit. on p. 15).

Fornwagner, Helena, Brit Grosskopf, Alexander Lauf, Vanessa Schöller, and Silvio Städter (2022a). *Gender versus sex: What drives behavior?* (Cit. on p. 85).

Fornwagner, Helena, Monika Pompeo, and Nina Serdarevic (2022b). “Choosing competition on behalf of someone else”. In: *Management Science* forthcoming (cit. on p. 85).

Forsythe, Robert, Joel L Horowitz, Nathan E Savin, and Martin Sefton (1994). “Fairness in simple bargaining experiments”. In: *Games and Economic Behavior* 6.3, pp. 347–369. DOI: 10.1006/game.1994.1021 (cit. on p. 183).

Fraley, Chris and Adrian E Raftery (2002). “Model-Based Clustering, Discriminant Analysis, and Density Estimation”. In: *Journal of the American Statistical Association* 97.458, pp. 611–631. DOI: 10.1198/016214502760047131 (cit. on p. 80).

Franzen, Axel and Sonja Pointner (2013). “The external validity of giving in the dictator game”. In: *Experimental Economics* 16.2, pp. 155–169. DOI: 10.1007/s10683-012-9337-5 (cit. on p. 183).

Galassi, Andrea, Marco Lippi, and Paolo Torroni (2020). “Attention in Natural Language Processing”. In: *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–18. DOI: 10.1109/TNNLS.2020.3019893 (cit. on p. 27).

Galizzi, Matteo M and Daniel Navarro-Martinez (2019). "On the external validity of social preference games: a systematic lab-field study". In: *Management Science* 65.3, pp. 976–1002. DOI: 10.1287/mnsc.2017.2908 (cit. on p. 183).

Garcia, R.C., J. Contreras, M. vanAkkeren, and J.B.C. Garcia (2005). "A GARCH Forecasting Model to Predict Day-Ahead Electricity Prices". In: *IEEE Transactions on Power Systems* 20.2, pp. 867–874. DOI: 10.1109/TPWRS.2005.846044 (cit. on p. 28).

Geldenhuys, Madelyn and Anita Bosch (2020). "A rasch adapted version of the 30-Item bem sex role inventory (BSRI)". In: *Journal of Personality Assessment* 102.3, pp. 428–439. DOI: 10.1080/00223891.2018.1527343 (cit. on pp. 88, 94).

Gianfreda, Angelica, Francesco Ravazzolo, and Luca Rossini (2020). "Comparing the forecasting performances of linear models for electricity prices with high RES penetration". In: *International Journal of Forecasting* 36.3, pp. 974–986. DOI: 10.1016/j.ijforecast.2019.11.002 (cit. on p. 22).

Gibbons, Jean D. and S. Chakraborti (1991). "Comparisons of the Mann-Whitney, Student's t, and Alternate t Tests for Means of Normal Distributions". In: *The Journal of Experimental Education* 59.3, pp. 258–267. DOI: 10.1080/00220973.1991.10806565 (cit. on p. 12).

Gilad, Dalia and Doron Kliger (2008). "Priming the risk attitudes of professionals in financial decision making". In: *Review of Finance* 12.3, pp. 567–586. DOI: 10.1093/rof/rfm034 (cit. on p. 183).

Glauner, Patrick, Manxing Du, Victor Paraschiv, Andrey Boytsov, Isabel Lopez Andrade, Jorge Meira, Petko Valtchev, and Radu State (2017). "The Top 10 Topics in Machine Learning Revisited: A Quantitative Meta-Study". In: *Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2017).* Bruges. arXiv: 1703.10121 (cit. on p. 23).

Gneezy, Uri, Kenneth L Leonard, and John A. List (2009). "Gender Differences in Competition: Evidence From a Matrilineal and a Patriarchal Society". In: *Econometrica* 77.5, pp. 1637–1664. DOI: 10.3982/ECTA6690 (cit. on pp. 77, 80, 84, 85).

Gneezy, Uri and Jan Potters (1997). "An experiment on risk taking and evaluation periods". In: *The Quarterly Journal of Economics* 112.2, pp. 631–645. DOI: 10.1162/003355397555217 (cit. on pp. 88, 182).

Gollier, Christian (2001). *The economics of risk and time.* MIT press (cit. on p. 182).

Gong, Binglin and Chun-Lei Yang (2012). "Gender differences in risk attitudes: Field experiments on the matrilineal Mosuo and the patriarchal Yi". In: *Journal of Economic Behavior & Organization* 83.1, pp. 59–65. DOI: 10.1016/j.jebo.2011.06.010 (cit. on p. 85).

Gonzalez, Jose Portela, Antonio Munoz San Roque, and Estrella Alonso Perez (2018). "Forecasting Functional Time Series with a New Hilbertian ARMAX Model: Application to Electricity Price Forecasting". In: *IEEE Transactions on Power Systems* 33.1, pp. 545–556. DOI: 10.1109/TPWRS.2017.2700287 (cit. on p. 23).

Goodfellow, Ian J., Yoshua Bengio, and Aaron Courville (2016). "Sequence Modeling : Recurrent and Recursive Nets". In: *Deep Learning*, pp. 324–365. DOI: 10.1533/9780857099440.59 (cit. on pp. 20, 101).

Graves, Alex (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Vol. 385. Studies in Computational Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-24797-2 (cit. on p. 20).

Greenland, Sander and James M. Robins (1985). "Estimation of a Common Effect Parameter from Sparse Follow-Up Data". In: *Biometrics* 41.1, p. 55. DOI: 10.2307/2530643 (cit. on p. 79).

Guiso, Luigi and Monica Paiella (2008). "Risk aversion, wealth, and background risk". In: *Journal of the European Economic Association* 6.6, pp. 1109–1150. DOI: 10.1162/JEEA.2008.6.6.1109 (cit. on pp. 85, 182).

Guiso, Luigi, Paola Sapienza, and Luigi Zingales (2018). "Time varying risk aversion". In: *Journal of Financial Economics* 128.3, pp. 403–421. DOI: 10.1016/j.jfineco.2018.02.007 (cit. on p. 183).

Haley, Kevin J and Daniel MT Fessler (2005). "Nobody's watching?: Subtle cues affect generosity in an anonymous economic game". In: *Evolution and Human Behavior* 26.3, pp. 245–256. DOI: 10.1016/j.evolhumbehav.2005.01.002 (cit. on p. 184).

Halko, Marja-Liisa and Lauri Sääksvuori (2017). "Competitive behavior, stress, and gender". In: *Journal of Economic Behavior & Organization* 141, pp. 96–109. DOI: 10.1016/j.jebo.2017.06.014 (cit. on p. 84).

Hardies, Kris, Diane Breesch, and Joël Branson (2013). "Gender differences in overconfidence and risk taking: Do self-selection and socialization matter?" In: *Economics Letters* 118.3, pp. 442–444. DOI: 10.1016/j.econlet.2012.12.004 (cit. on p. 182).

Harrer, M, P Cuijpers, TA Furukawa, and DD Ebert (2019). "Doing meta-analysis in R: A hands-on guide". In: *PROTECT Lab Erlangen* (cit. on p. 80).

Hartigan, J. A. and M. A. Wong (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". In: *Applied Statistics* 28.1, pp. 100–108. DOI: 10.2307/2346830 (cit. on p. 80).

He, Joyce C, Sonia K Kang, and Nicola Lacetera (2021). "Opt-out choice framing attenuates gender differences in the decision to compete in the laboratory and in the field". In: *Proceedings of the National Academy of Sciences* 118.42. DOI: 10.1073/pnas.2108337118 (cit. on p. 85).

Healy, Andrew and Jennifer Pate (2011). "Can Teams Help to Close the Gender Competition Gap?" In: *The Economic Journal* 121.555, pp. 1192–1204. DOI: 10.1111/j.1468-0297.2010.02409.x (cit. on p. 84).

Higgins, Julian P. T., James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch, eds. (2019). *Cochrane handbook for systematic reviews of interventions*. Second edition. Cochrane Book Series. Hoboken, NJ: Wiley-Blackwell (cit. on p. 78).

Hinderks, W.J. and A. Wagner (2019). "Factor models in the German electricity market: Stylized facts, seasonality, and calibration". In: *Energy Economics*. DOI: `10.1016/J.ENECO.2019.03.024` (cit. on p. 35).

Hirth, Lion and Inka Ziegenhagen (2015). "Balancing power and variable renewables: Three links". In: *Renewable and Sustainable Energy Reviews* 50, pp. 1035–1051. DOI: `10.1016/j.rser.2015.04.180` (cit. on p. 24).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. DOI: `10.1162/neco.1997.9.8.1735` (cit. on pp. 25, 31).

Hogarth, Robin M, Natalia Karelaia, and Carlos Andrés Trujillo (2012). "When should I quit? Gender differences in exiting competitions". In: *Journal of Economic Behavior & Organization* 83.1, pp. 136–150. DOI: `10.1016/j.jebo.2011.06.021` (cit. on p. 77).

Holt, Charles A and Susan K Laury (2002). "Risk aversion and incentive effects". In: *American Economic Review* 92.5, pp. 1644–1655. DOI: `10.1257/000282802762024700` (cit. on p. 182).

Hong, Tao, Pierre Pinson, and Shu Fan (2014). "Global Energy Forecasting Competition 2012". In: *International Journal of Forecasting* 30.2, pp. 357–363. DOI: `10.1016/J.IJFORECAST.2013.07.001` (cit. on p. 22).

Hong, Tao, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J. Hyndman (2016). "Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond". In: *International Journal of Forecasting* 32.3, pp. 896–913. DOI: `10.1016/j.ijforecast.2016.02.001` (cit. on p. 22).

Hong, Tao, Jingrui Xie, and Jonathan Black (2019). "Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting". In: *International Journal of Forecasting* 35.4, pp. 1389–1399. DOI: `10.1016/J.IJFORECAST.2019.02.006` (cit. on p. 22).

Hörmann, Siegfried and Piotr Kokoszka (2012). "Functional Time Series". In: *Handbook of Statistics*. Vol. 30. Elsevier, pp. 157–186. DOI: `10.1016/B978-0-444-53858-1.00007-7` (cit. on p. 29).

Horváth, Lajos, Piotr Kokoszka, and Gregory Rice (2014). "Testing stationarity of functional time series". In: *Journal of Econometrics* 179.1, pp. 66–82. DOI: `10.1016/j.jeconom.2013.11.002` (cit. on p. 40).

Hukkanen, Petri and Matti Keloharju (2019). "Initial Offer Precision and M&A Outcomes: Initial Offer Precision and M&A Outcomes". In: *Financial Management* 48.1, pp. 291–310. DOI: `10.1111/fima.12229` (cit. on pp. 53, 57).

Hyde, Janet Shibley, Rebecca S Bigler, Daphna Joel, Charlotte Chucky Tate, and Sari M van Anders (2019). "The future of sex and gender in psychology: Five challenges to the gender binary." In: *American Psychologist* 74.2, pp. 171–193. DOI: `10.1037/amp0000307` (cit. on p. 93).

Hyndman, Rob J and Han Lin Shang (2009). "Functional time series forecasting". In: *Journal of the Korean Statistical Society* 38.3, pp. 199–211. DOI: 10.1016/j.jkss.2009.06.002 (cit. on pp. 30, 31).

Hyndman, Rob J. and Yeasmin Khandakar (2008). "Automatic Time Series Forecasting: The **forecast** Package for *R*". In: *Journal of Statistical Software* 27.3. DOI: 10.18637/jss.v027.i03 (cit. on p. 43).

Hyndman, Rob J. and Md. Shahid Ullah (2007). "Robust forecasting of mortality and fertility rates: A functional data approach". In: *Computational Statistics & Data Analysis* 51.10, pp. 4942–4956. DOI: 10.1016/j.csda.2006.07.028 (cit. on pp. 30, 31).

IntHout, Joanna, John PA Ioannidis, and George F Borm (2014). "The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method". In: *BMC Medical Research Methodology* 14.1, p. 25. DOI: 10.1186/1471-2288-14-25 (cit. on p. 79).

Isoni, Andrea, Anders Poulsen, Robert Sugden, and Kei Tsutsui (2013). "Focal points in tacit bargaining problems: Experimental evidence". In: *European Economic Review* 59, pp. 167–188. DOI: 10.1016/j.euroecorev.2012.12.005 (cit. on p. 55).

Isoni, Andrea, Anders Poulsen, Robert Sugden, and Kei Tsutsui (2019). "Focal points and payoff information in tacit bargaining". In: *Games and Economic Behavior* 114, pp. 193–214. DOI: 10.1016/j.geb.2019.01.008 (cit. on p. 55).

Janiszewski, Chris and Dan Uy (2008). "Precision of the Anchor Influences the Amount of Adjustment". In: *Psychological Science* 19.2, pp. 121–127. DOI: 10.1111/j.1467-9280.2008.02057.x (cit. on pp. 53, 57).

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler (1986). "Fairness and the assumptions of economics". In: *Journal of Business* 59.4, pp. 285–300 (cit. on pp. 88, 183).

Kalchbrenner, Nal and Phil Blunsom (2013). "Recurrent Continuous Translation Models". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1700–1709. URL: https://www.aclweb.org/anthology/D13-1176 (cit. on p. 20).

Kamas, Linda and Anne Preston (2012). "The importance of being confident; gender, career choice, and willingness to compete". In: *Journal of Economic Behavior & Organization* 83.1, pp. 82–97. DOI: 10.1016/j.jebo.2011.06.013 (cit. on p. 84).

Kamas, Linda and Anne Preston (2015). "Can social preferences explain gender differences in economic behavior?" In: *Journal of Economic Behavior & Organization* 116, pp. 525–539. DOI: 10.1016/j.jebo.2015.05.017 (cit. on p. 77).

Kastlunger, Barbara, Stefan G Dressler, Erich Kirchler, Luigi Mittone, and Martin Voracek (2010). "Sex differences in tax compliance: Differentiating between demographic sex, gender-role orientation, and prenatal masculinization (2D: 4D)". In: *Journal of Economic Psychology* 31.4, pp. 542–552. DOI: 10.1016/j.joep.2010.03.015 (cit. on p. 93).

Keaveney, Alexis, Ellen Peters, and Baldwin Way (2020). "Effects of acetaminophen on risk taking". In: *Social Cognitive and Affective Neuroscience* 15.7, pp. 725–732. DOI: 10.1093/scan/nsaa108 (cit. on p. 182).

Kessel, Dany, Johanna Mollerstrom, and Roel van Veldhuizen (2021). "Can simple advice eliminate the gender gap in willingness to compete?" In: *European Economic Review* 138, p. 103777. DOI: 10.1016/j.euroecorev.2021.103777 (cit. on p. 82).

Khachatryan, Karen, Anna Dreber, Emma von Essen, and Eva Ranehill (2015). "Gender and preferences at a young age: Evidence from Armenia". In: *Journal of Economic Behavior & Organization* 118, pp. 318–332. DOI: 10.1016/j.jebo.2015.02.021 (cit. on p. 84).

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representation 2015*. arXiv: 1412.6980 (cit. on p. 42).

Klæboe, Gro, Anders Lund Eriksrud, and Stein-Erik Fleten (2015). "Benchmarking time series based forecasting models for electricity balancing market prices". In: *Energy Systems* 6.1, pp. 43–61. DOI: 10.1007/s12667-013-0103-3 (cit. on pp. 23, 39).

Klege, Rebecca Afua, Martine Visser, Manuel F. Barron A, and Rowan P. Clarke (2021). "Competition and gender in the lab vs field: Experiments from off-grid renewable energy entrepreneurs in Rural Rwanda". In: *Journal of Behavioral and Experimental Economics* 91, p. 101662. DOI: 10.1016/j.socec.2021.101662 (cit. on pp. 78, 84).

Klepsch, J., C. Klüppelberg, and T. Wei (2017). "Prediction of functional ARMA processes with an application to traffic data". In: *Econometrics and Statistics* 1, pp. 128–149. DOI: 10.1016/j.ecosta.2016.10.009 (cit. on pp. 30, 31).

Klinowski, David (2019). "Selection into self-improvement and competition pay: Gender, stereotypes, and earnings volatility". In: *Journal of Economic Behavior & Organization* 158, pp. 128–146. DOI: 10.1016/j.jebo.2018.11.014 (cit. on p. 84).

Knapp, Guido and Joachim Hartung (2003). "Improved tests for a random effects meta-regression with a single covariate". In: *Statistics in Medicine* 22.17, pp. 2693–2710. DOI: 10.1002/sim.1482 (cit. on p. 79).

Knief, Ulrich and Wolfgang Forstmeier (2021). "Violating the normality assumption may be the lesser of two evils". In: *Behavior Research Methods*. DOI: 10.3758/s13428-021-01587-5 (cit. on pp. 12, 16).

Kokoszka, Piotr and Matthew Reimherr (2013). "Determining the order of the functional autoregressive model". In: *Journal of Time Series Analysis* 34.1, pp. 116–129. DOI: 10.1111/j.1467-9892.2012.00816.x (cit. on p. 30).

Kokoszka, Piotr and Han Lin Shang (2017). *Stationarity tests and a new prediction method for functional time series* (cit. on p. 40).

König-Kersting, Christian and Stefan T Trautmann (2018). "Countercyclical risk aversion: Beyond financial professionals". In: *Journal of Behavioral and Experimental Finance* 18, pp. 94–101. DOI: 10.1016/j.jbef.2018.03.001 (cit. on p. 183).

Kozee, Holly B, Tracy L Tylka, and L Andrew Bauerband (2012). “Measuring transgender individuals’ comfort with gender identity and appearance: Development and validation of the Transgender Congruence Scale”. In: *Psychology of Women Quarterly* 36.2, pp. 179–196. DOI: 10.1177/0361684312442161 (cit. on p. 88).

Lacetera, Nicola, Devin G Pope, and Justin R Sydnor (2012). “Heuristic thinking and limited attention in the car market”. In: *American Economic Review* 102.5, pp. 2206–2236. DOI: 10.1257/aer.102.5.2206 (cit. on pp. 53, 55).

Lago, Jesus, Fjo De Ridder, and Bart De Schutter (2018). “Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms”. In: *Applied Energy* 221, pp. 386–405. DOI: 10.1016/j.apenergy.2018.02.069 (cit. on pp. 23, 27).

Lauf, Alexander and Lars Schlereth (2022). *Round-number Effects in Bargaining: Bias vs. Focal Point* (cit. on p. 53).

LeCun, Yann (1989). *Generalization and network design strategies*. Technical Report CRG-TR-89-4. University of Toronto (cit. on pp. 30, 31).

Lee, Soohyung, Muriel Niederle, and Namwook Kang (2014). “Do single-sex schools make girls more competitive?” In: *Economics Letters* 124.3, pp. 474–477. DOI: 10.1016/j.econlet.2014.07.001 (cit. on p. 84).

Leibbrandt, Andreas and John A List (2015). “Do women avoid salary negotiations? Evidence from a large-scale natural field experiment”. In: *Management Science* 61.9, pp. 2016–2024. DOI: 10.1287/mnsc.2014.1994 (cit. on p. 77).

Leibbrandt, Andreas, Liang Choon Wang, and Cordelia Foo (2018). “Gender quotas, competitions, and peer review: Experimental evidence on the backlash against women”. In: *Management Science* 64.8, pp. 3501–3516. DOI: 10.1287/mnsc.2017.2772 (cit. on p. 82).

Lemaster, Philip and JoNell Strough (2014). “Beyond Mars and Venus: Understanding gender differences in financial risk tolerance”. In: *Journal of Economic Psychology* 42, pp. 148–160. DOI: 10.1016/j.joep.2013.11.001 (cit. on p. 93).

Li, Lisha, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar (2018). “Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization”. In: *Journal of Machine Learning Research* 18, pp. 1–52 (cit. on p. 42).

Liebl, Dominik and Stefan Rameseder (2019). “Partially Observed Functional Data: The Case of Systematically Missing Parts”. In: *Computational Statistics and Data Analysis* 131, pp. 104–115. arXiv: 1711.07715v3 (cit. on p. 23).

Lim, Bryan and Stefan Zohren (2021). “Time-series forecasting with deep learning: a survey”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2194, p. 20200209. DOI: 10.1098/rsta.2020.0209 (cit. on pp. 23, 27).

Lin, Po-Hsuan, Alexander L. Brown, Taisuke Imai, Joseph Tao-yi Wang, Stephanie W. Wang, and Colin F. Camerer (2020). “Evidence of general economic principles of bargaining and trade from 2,000 classroom experiments". In: *Nature Human Behaviour* 4.9, pp. 917–927. DOI: 10.1038/s41562-020-0916-8 (cit. on p. 56).

Liu, Elaine M and Sharon Xuejing Zuo (2019). "Measuring the impact of interaction between children of a matrilineal and a patriarchal culture on gender differences in risk aversion". In: *Proceedings of the National Academy of Sciences* 116.14, pp. 6713–6719. DOI: 10.1073/pnas.1808336116 (cit. on p. 85).

Liu, Heping and Jing Shi (2013). "Applying ARMA–GARCH approaches to forecasting short-term electricity prices". In: *Energy Economics* 37, pp. 152–166. DOI: 10.1016/j.eneco.2013.02.006 (cit. on p. 28).

Loschelder, David D., Johannes Stuppi, and Roman Trötschel (2014). ""€14,875?!": Precision Boosts the Anchoring Potency of First Offers". In: *Social Psychological and Personality Science* 5.4, pp. 491–499. DOI: 10.1177/1948550613499942 (cit. on pp. 53, 57).

Lucas, Alexandre, Konstantinos Pegios, Evangelos Kotsakis, and Dan Clarke (2020). "Price Forecasting for the Balancing Energy Market Using Machine-Learning Regression". In: *Energies* 13.20, p. 5420. DOI: 10.3390/en13205420 (cit. on p. 24).

Mantel, Nathan and William Haenszel (1959). "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease". In: 22.4, pp. 719–748. DOI: 10.1093/jnci/22.4.719 (cit. on p. 79).

Marianne, Bertrand (2011). "New perspectives on gender". In: *Handbook of labor economics*. Vol. 4. Elsevier, pp. 1543–1590 (cit. on p. 94).

Mason, Malia F., Alice J. Lee, Elizabeth A. Wiley, and Daniel R. Ames (2013). "Precise offers are potent anchors: Conciliatory counteroffers and attributions of knowledge in negotiations". In: *Journal of Experimental Social Psychology* 49.4, pp. 759–763. DOI: 10.1016/j.jesp.2013.02.012 (cit. on pp. 53, 57).

Mayr, Ulrich, Dave Wozniak, Casey Davidson, David Kuhns, and William T. Harbaugh (2012). "Competitiveness across the life span: The feisty fifties." In: *Psychology and Aging* 27.2, pp. 278–285. DOI: 10.1037/a0025655 (cit. on p. 84).

Mazzi, Nicolo, Jalal Kazempour, and Pierre Pinson (2018). "Price-Taker Offering Strategy in Electricity Pay-as-Bid Markets". In: *IEEE Transactions on Power Systems* 33.2, pp. 2175–2183. DOI: 10.1109/TPWRS.2017.2737322 (cit. on p. 23).

McDowall, David, Richard McCleary, and Bradley J. Bartos (2019). *Interrupted time series analysis*. New York, NY: Oxford University Press. 180 pp. (cit. on p. 107).

McKay, Ryan, Charles Efferson, Harvey Whitehouse, and Ernst Fehr (2011). "Wrath of God: Religious primes and punishment". In: *Proceedings of the Royal Society B: Biological Sciences* 278.1713, pp. 1858–1863. DOI: 10.1098/rspb.2010.2125 (cit. on p. 184).

Mehta, Judith, Chris Starmer, and Robert Sugden (1994a). "Focal points in pure coordination games: An experimental investigation". In: *Theory and Decision* 36.2, pp. 163–185. DOI: 10.1007/BF01079211 (cit. on p. 54).

Mehta, Judith, Chris Starmer, and Robert Sugden (1994b). "The Nature of Salience: An Experimental Investigation of Pure Coordination Games". In: *American Economic Review* 84.3, pp. 658–673 (cit. on p. 54).

Meier-Pesti, Katja and Elfriede Penz (2008). "Sex or gender? Expanding the sex-based view by introducing masculinity and femininity as predictors of financial risk taking". In: *Journal of Economic Psychology* 29.2, pp. 180–196. DOI: 10.1016/j.joep.2007.05.002 (cit. on pp. 93, 183).

Meyer, Mary C. (2008). "Inference using shape-restricted regression splines". In: *The Annals of Applied Statistics* 2.3, pp. 1013–1033. DOI: 10.1214/08-AOAS167 (cit. on p. 11).

Meyer, Mary C. (2012). "Constrained penalized splines". In: *Canadian Journal of Statistics* 40.1, pp. 190–206. DOI: 10.1002/cjs.10137 (cit. on pp. 11, 39).

Moll, Jorge, Frank Krueger, Roland Zahn, Matteo Pardini, Ricardo de Oliveira-Souza, and Jordan Grafman (2006). "Human fronto–mesolimbic networks guide decisions about charitable donation". In: *Proceedings of the National Academy of Sciences* 103.42, pp. 15623–15628. DOI: 10.1073/pnas.0604475103 (cit. on p. 85).

Müller, Julia and Christiane Schwieren (2012). "Can personality explain what is underlying women's unwillingness to compete?" In: *Journal of Economic Psychology*, pp. 448–460. DOI: 10.1016/j.joep.2011.12.005 (cit. on p. 84).

Murthy, D. N. P., M. Xie, and Renyan Jiang (2004). *Weibull models*. Wiley Series in Probability and Statistics. Hoboken, N.J: J. Wiley. 383 pp. (cit. on p. 14).

Müsgens, Felix, Axel Ockenfels, and Markus Peek (2014). "Economics and design of balancing power markets in Germany". In: *International Journal of Electrical Power & Energy Systems* 55, pp. 392–401. DOI: 10.1016/j.ijepes.2013.09.020 (cit. on p. 23).

Newell, Ben R and Brad Shaw (2017). "Priming risky choice: Do risk preferences need inferences?" In: *Journal of Behavioral Decision Making* 30.2, pp. 332–346. DOI: 10.1002/bdm.1945 (cit. on p. 183).

Niederle, Muriel (2016). "Gender". In: *Handbook of Experimental Economics*. Ed. by John Kagel and Alvin E Roth. 2nd ed. Princeton: Princeton University Press, pp. 481–562 (cit. on pp. 70, 78).

Niederle, Muriel (2017). "A gender agenda: A progress report on competitiveness". In: *American Economic Review* 107.5, pp. 115–119. DOI: 10.1257/aer.p20171066 (cit. on p. 77).

Niederle, Muriel, Carmit Segal, and Lise Vesterlund (2013). "How costly is diversity? Affirmative action in light of gender differences in competitiveness". In: *Management Science* 59.1, pp. 1–16. DOI: 10.1287/mnsc.1120.1602 (cit. on p. 84).

Niederle, Muriel and Lise Vesterlund (2007). "Do Women Shy Away From Competition? Do Men Compete Too Much?" In: *The Quarterly Journal of Economics* 122.3, pp. 1067–1101. DOI: 10.1162/qjec.122.3.1067 (cit. on pp. 77, 79, 82, 84, 85).

Nowotarski, Jakub and Rafał Weron (2018). “Recent advances in electricity price forecasting: A review of probabilistic forecasting”. In: *Renewable and Sustainable Energy Reviews* 81 (September 2016), pp. 1548–1568. DOI: 10.1016/j.rser.2017.05.234 (cit. on p. 22).

O’Connor, Patrick D. T. and Andre Kleyner (2011). *Practical Reliability Engineering: O’Connor/Practical Reliability Engineering*. Chichester, UK: John Wiley & Sons, Ltd. DOI: 10.1002/9781119961260 (cit. on p. 14).

O’Sullivan, Finbarr (1986). “A statistical perspective on ill-posed inverse problems”. In: *Statistical science*, pp. 502–518 (cit. on p. 10).

Ocker, Fabian, Sebastian Braun, and Christian Will (2016). “Design of European balancing power markets”. In: *2016 13th International Conference on the European Energy Market (EEM)*. IEEE, pp. 1–6. DOI: 10.1109/EEM.2016.7521193 (cit. on p. 23).

Ocker, Fabian, Karl-Martin Ehrhart, and Matej Belica (2018). “Harmonization of the European balancing power auction: A game-theoretical and empirical investigation”. In: *Energy Economics* 73, pp. 194–211. DOI: 10.1016/J.ENECO.2018.05.003 (cit. on pp. 19, 23, 24).

Olsson, Magnus and Lennart Soder (2008). “Modeling Real-Time Balancing Power Market Prices Using Combined SARIMA and Markov Processes”. In: *IEEE Transactions on Power Systems* 23.2, pp. 443–450. DOI: 10.1109/TPWRS.2008.920046 (cit. on p. 23).

Online (2022a). *As Predicted: On the robustness of gender differences in economic behavior*. URL: https://aspredicted.org/blind.php?x=DCL_1VB (visited on 03/16/2022) (cit. on p. 149).

Online (2022b). *Best Offer Sequential Bargaining*. URL: https://www.nber.org/research/data/best-offer-sequential-bargaining (visited on 03/16/2022) (cit. on p. 113).

Online (2022c). *Definition - Sex*. Oxford English Dictionary. URL: https://www.oed.com/view/Entry/176989 (visited on 03/16/2022) (cit. on p. 85).

Online (2022d). *Keras documentation: Keras Applications*. URL: https://keras.io/api/applications/ (visited on 03/16/2022) (cit. on p. 51).

Online (2022e). *Keras documentation: KerasTuner*. URL: https://keras.io/keras_tuner/ (visited on 03/16/2022) (cit. on p. 42).

Online (2022f). *Keras documentation: Why choose Keras?* URL: https://keras.io/why_keras/ (visited on 03/16/2022) (cit. on p. 41).

Online (2022g). *OSF: Gender versus sex: What drives behavior?* URL: https://osf.io/tyzjh/?view_only=66a8abca5f6a4aeead68f6fef19a0ee9 (visited on 03/16/2022) (cit. on p. 185).

Online (2022h). *OSF: Neural Functional Time Series Forecasting - Application to the German Balancing Market*. URL: https://osf.io/4fz9c/?view_only=d45e0e1b46eb470095129a38a249daf5 (visited on 03/16/2022) (cit. on p. 35).

Online (2022i). *OSF: On gender differences in competitiveness*. URL: https://osf.io/ymxgj/?view_only=41e01758de7c4ba0ab7e9b186f4f69c1 (visited on 03/18/2022) (cit. on p. 78).

Online (2022j). *Prolific · Quickly find research participants you can trust.* URL: https://www.prolific.co/ (visited on 03/17/2022) (cit. on p. 87).

Online (2022k). *Shiny: K-order B-spline basis functions.* URL: https://alexander-lauf.shinyapps.io/B-Spline_basis_functions/ (visited on 03/16/2022) (cit. on p. 9).

Palm, F.C. (1996). "7 GARCH models of volatility". In: *Handbook of Statistics.* Vol. 14. Elsevier, pp. 209–240. DOI: 10.1016/S0169-7161(96)14009-8 (cit. on p. 29).

Parravano, Melanie and Odile Poulsen (2015). "Stake size and the power of focal points in coordination games: Experimental evidence". In: *Games and Economic Behavior* 94, pp. 191–199. DOI: 10.1016/j.geb.2015.05.001 (cit. on p. 54).

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio (2013). "On the difficulty of training recurrent neural networks". In: *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)* (cit. on p. 25).

Perperoglou, Aris, Willi Sauerbrei, Michal Abrahamowicz, and Matthias Schmid (2019). "A review of spline function procedures in R". In: *BMC Medical Research Methodology* 19.1, p. 46. DOI: 10.1186/s12874-019-0666-3 (cit. on p. 9).

Pope, Devin and Uri Simonsohn (2011). "Round Numbers as Goals: Evidence From Baseball, SAT Takers, and the Lab". In: *Psychological Science* 22.1, pp. 71–79. DOI: 10.1177/0956797610391098 (cit. on p. 55).

Pope, Devin G., Jaren C. Pope, and Justin R. Sydnor (2015). "Focal points and bargaining in housing markets". In: *Games and Economic Behavior* 93, pp. 89–107. DOI: 10.1016/j.geb.2015.07.002 (cit. on pp. 53, 56).

Poplavskaya, Ksenia and Laurens de Vries (2019). "Distributed energy resources and the organized balancing market: A symbiosis yet? Case of three European balancing markets". In: *Energy Policy* 126, pp. 264–276. DOI: 10.1016/j.enpol.2018.11.009 (cit. on pp. 23, 24).

Poplavskaya, Ksenia, Jesus Lago, and Laurens de Vries (2020). "Effect of market design on strategic bidding behavior: Model-based analysis of European electricity balancing markets". In: *Applied Energy* 270, p. 115130. DOI: 10.1016/j.apenergy.2020.115130 (cit. on p. 23).

Price, Curtis R. (2012). "Gender, Competition, and Managerial Decisions". In: *Management Science* 58.1, pp. 114–122. DOI: 10.1287/mnsc.1110.1384 (cit. on p. 84).

Ramsay, J. O. (1982). "When the data are functions". In: *Psychometrika* 47.4, pp. 379–396. DOI: 10.1007/BF02293704 (cit. on p. 5).

Ramsay, J. O. and C. J. Dalzell (1991). "Some Tools for Functional Data Analysis". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3, pp. 539–561. DOI: 10.1111/j.2517-6161.1991.tb01844.x (cit. on p. 5).

Ramsay, J. O., Giles Hooker, and Spencer Graves (2009). *Functional data analysis with R and MATLAB.* Use R! Dordrecht ; New York: Springer. 207 pp. (cit. on p. 39).

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis*. 2nd ed. Springer Series in Statistics. New York: Springer. 426 pp. (cit. on pp. 6, 10).

Ranehill, Eva, Niklas Zethraeus, Liselott Blomberg, Bo von Schoultz, Angelica Lindén Hirschberg, Magnus Johannesson, and Anna Dreber (2018). "Hormonal contraceptives do not impact economic preferences: Evidence from a randomized trial". In: *Management Science* 64.10, pp. 4515–4532. DOI: 10.1287/mnsc.2017.2844 (cit. on pp. 83, 85, 182).

Repetto, Luca and Alex Solís (2019). "The Price of Inattention: Evidence from the Swedish Housing Market". In: *Journal of the European Economic Association*, jvz065. DOI: 10.1093/jeea/jvz065 (cit. on p. 53).

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales (2015). *Taste for competition and the gender gap among young business professionals*. Working paper 21695. National Bureau of Economic Research (cit. on p. 77).

Reuben, Ernesto, Matthew Wiswall, and Basit Zafar (2017). "Preferences and Biases in Educational Choices and Labour Market Expectations: Shrinking the Black Box of Gender". In: *The Economic Journal* 127.604, pp. 2153–2186. DOI: 10.1111/ecoj.12350 (cit. on pp. 77, 84).

Reuter, Martin, Clemens Frenzel, Nora T Walter, Sebastian Markett, and Christian Montag (2011). "Investigating the genetic basis of altruism: The role of the COMT Val158Met polymorphism". In: *Social Cognitive and Affective Neuroscience* 6.5, pp. 662–668. DOI: 10.1093/scan/nsq083 (cit. on pp. 85, 184).

Rigdon, Mary, Keiko Ishii, Motoki Watabe, and Shinobu Kitayama (2009). "Minimal social cues in the dictator game". In: *Journal of Economic Psychology* 30.3, pp. 358–367. DOI: 10.1016/j.joep.2009.02.002 (cit. on p. 184).

Rosch, Eleanor (1975). "Cognitive reference points". In: *Cognitive Psychology* 7.4, pp. 532–547. DOI: 10.1016/0010-0285(75)90021-3 (cit. on p. 55).

Rudman, Laurie A and Julie E Phelan (2010). "The effect of priming gender roles on women's implicit gender beliefs and career aspirations". In: *Social Psychology* 41.3, pp. 192–202. DOI: 10.1027/1864-9335/a000027 (cit. on p. 86).

Rumelhart, David E, Geoffrey E Hintont, and Ronald J Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323.6088, pp. 533–536. DOI: 10.1038/323533a0 (cit. on pp. 20, 41).

Saccardo, Silvia, Aniela Pietrasz, and Uri Gneezy (2018). "On the size of the gender difference in competitiveness". In: *Management Science* 64.4, pp. 1541–1554 (cit. on p. 77).

Samak, Anya C. (2013). "Is there a gender gap in preschoolers' competitiveness? An experiment in the U.S". In: *Journal of Economic Behavior & Organization* 92, pp. 22–31. DOI: 10.1016/j.jebo.2013.04.014 (cit. on p. 84).

Sapienza, Paola, Luigi Zingales, and Dario Maestripieri (2009). "Gender differences in financial risk aversion and career choices are affected by testosterone". In: *Proceedings* *of the National Academy of Sciences* 106.36, pp. 15268–15273. DOI: 10.1073/pnas.0907352106 (cit. on p. 85).

Schelling, Thomas (1960). *The Strategy of Conflict*. Harvard University Press (cit. on p. 54).

Schwarzer, Guido, James R. Carpenter, and Gerta Rücker (2015). *Meta-Analysis with R*. Use R! Cham: Springer International Publishing. DOI: 10.1007/978-3-319-21416-0 (cit. on pp. 78, 79).

Shah, Ismail and Francesco Lisi (2020). "Forecasting of electricity price through a functional prediction of sale and purchase curves". In: *Journal of Forecasting* 39.2, pp. 242–259. DOI: 10.1002/for.2624 (cit. on pp. 22, 52, 53).

Shang, Han Lin (2013). "Ftsa : An R Package for Analyzing Functional Time Series". In: *The R Journal Vol* 5 (June), pp. 64–72. DOI: 10.32614/RJ-2013-006 (cit. on pp. 30, 31).

Shariff, Azim F and Ara Norenzayan (2007). "God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game". In: *Psychological Science* 18.9, pp. 803–809. DOI: 10.1111/j.1467-9280.2007.01983.x (cit. on p. 184).

Shurchkov, Olga (2012). "Under pressure: Gender differences in output quality and quantity under competition and time constraints". In: *Journal of the European Economic Association* 10.5, pp. 1189–1213. DOI: 10.1111/j.1542-4774.2012.01084.x (cit. on p. 84).

Shurchkov, Olga and Catherine C Eckel (2018). *Gender differences in behavioral traits and labor market outcomes*. Oxford, UK: Oxford University Press (cit. on p. 77).

Sobkowicz, Pawel, Mike Thelwall, Kevan Buckley, Georgios Paltoglou, and Antoni Sobkowicz (2013). "Lognormal distributions of user post lengths in Internet discussions - a consequence of the Weber-Fechner law?" In: *EPJ Data Science* 2.1, p. 2. DOI: 10.1140/epjds14 (cit. on p. 14).

Steele, Jennifer R and Nalini Ambady (2006). ""Math is Hard!" The effect of gender priming on women's attitudes". In: *Journal of Experimental Social Psychology* 42.4, pp. 428–436. DOI: 10.1016/j.jesp.2005.06.003 (cit. on p. 86).

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to Sequence Learning with Neural Networks". In: pp. 1–9. DOI: 10.1007/s10107-014-0839-0 (cit. on p. 20).

Sutter, Matthias and Daniela Glätzle-Rützler (2015). "Gender Differences in the Willingness to Compete Emerge Early in Life and Persist". In: *Management Science* 61.10, pp. 2339–2354. DOI: 10.1287/mnsc.2014.1981 (cit. on pp. 77, 84, 85).

Sutter, Matthias, Daniela Glätzle-Rützler, Loukas Balafoutas, and Simon Czermak (2016). "Cancelling out early age gender differences in competition: an analysis of policy interventions". In: *Experimental Economics* 19.2, pp. 412–432. DOI: 10.1007/s10683-015-9447-y (cit. on p. 84).

Sutter, Matthias, Martin G Kocher, Daniela Glätzle-Rützler, and Stefan T Trautmann (2013). "Impatience and uncertainty: Experimental decisions predict adolescents' field behavior". In: *American Economic Review* 103.1, pp. 510–31. DOI: 10.1257/aer.103.1.510 (cit. on p. 182).

Thöni, Christian and Stefan Volk (2021). "Converging evidence for greater male variability in time, risk, and social preferences". In: *Proceedings of the National Academy of Sciences* 118.23, e2026112118. DOI: 10.1073/pnas.2026112118 (cit. on p. 85).

Van Anders, Sari M, Jeffrey Steiger, and Katherine L Goldey (2015). "Effects of gendered behavior on testosterone in women and men". In: *Proceedings of the National Academy of Sciences* 112.45, pp. 13805–13810. DOI: 10.1073/pnas.1509591112 (cit. on p. 85).

Van der Veen, Reinier A.C. and Rudi A. Hakvoort (2016). "The electricity balancing market: Exploring the design challenge". In: *Utilities Policy* 43.2, pp. 186–194. DOI: 10.1016/j.jup.2016.10.008 (cit. on pp. 19, 23, 33).

VandenBos, Gary R (2007). *APA dictionary of psychology.* American Psychological Association (cit. on p. 86).

Vandezande, Leen, Leonardo Meeus, Ronnie Belmans, Marcelo Saguan, and Jean Michel Glachant (2010). "Well-functioning balancing markets: A prerequisite for wind power integration". In: *Energy Policy* 38.7, pp. 3146–3154. DOI: 10.1016/j.enpol.2009.07.034 (cit. on p. 23).

Villeval, Marie Claire (2012). "Ready, steady, compete". In: *Science (New York, N.Y.)* 335.6068, pp. 544–545. DOI: 10.1126/science.1218000 (cit. on p. 85).

Von Gaudecker, Hans-Martin, Arthur Van Soest, and Erik Wengstrom (2011). "Heterogeneity in risky choice behavior in a broad population". In: *American Economic Review* 101.2, pp. 664–94. DOI: 10.1257/aer.101.2.664 (cit. on pp. 85, 182).

Wadhwa, Monica and Kuangjie Zhang (2019). "When numbers make you feel: Impact of round versus precise numbers on preventive health behaviors". In: *Organizational Behavior and Human Decision Processes* 150, pp. 101–111. DOI: 10.1016/j.obhdp.2018.08.005 (cit. on p. 55).

Wang, Jane-Ling, Jeng-Min Chiou, and Hans-Georg Müller (2016). "Functional Data Analysis". In: *Annual Review of Statistics and Its Application* 3, pp. 257–295. DOI: 10.1146/annurev-statistics-041715-033624 (cit. on p. 20).

Weron, Rafał (2006). *Modeling and Forecasting Electricity Loads and Prices.* John Wiley & Sons Ltd (cit. on p. 23).

Weron, Rafał (2014). "Electricity price forecasting: A review of the state-of-the-art with a look into the future". In: *International Journal of Forecasting* 30.4, pp. 1030–1081. DOI: 10.1016/j.ijforecast.2014.08.008 (cit. on p. 22).

Weron, Rafał and Adam Misiorek (2008). "Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models". In: *International Journal of Forecasting* 24, pp. 744–763. DOI: 10.1016/j.ijforecast.2008.08.004 (cit. on p. 28).

Weron, Rafał and Florian Ziel (2018). "Electricity price forecasting". In: *Handbook of Energy Economics.* Ed. by Ugur Soytas and Ramazan Sari. 1st. Routledge (cit. on p. 23).

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Pedersen, Evan Miller, Stephan Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani (2019). "Welcome to the Tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686 (cit. on p. 35).

Williams, Ronald J. and David Zipser (1989). "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks". In: *Neural Computation* 1.2, pp. 270–280. DOI: 10.1162/neco.1989.1.2.270 (cit. on p. 25).

Wojna, Zbigniew, Vittorio Ferrari, Sergio Guadarrama, Nathan Silberman, Liang-Chieh Chen, Alireza Fathi, and Jasper Uijlings (2019). "The Devil is in the Decoder: Classification, Regression and GANs". In: *International Journal of Computer Vision*. arXiv: 1707.05847v3 (cit. on p. 27).

Wozniak, David, William T Harbaugh, and Ulrich Mayr (2016). *The effect of feedback on gender differences in competitive choices.* Working paper 1976073. SSRN. DOI: 10.2139/ssrn.1976073 (cit. on p. 82).

Wozniak, David, William T. Harbaugh, and Ulrich Mayr (2014). "The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices". In: *Journal of Labor Economics* 32.1, pp. 161–198. DOI: 10.1086/673324 (cit. on pp. 83, 84).

Yan, Dengfeng and Jorge Pena-Marin (2017). "Round Off the Bargaining: The Effects of Offer Roundness on Willingness to Accept". In: *Journal of Consumer Research* 44.2, pp. 381–395. DOI: 10.1093/jcr/ucx046 (cit. on pp. 53, 55).

Zak, Paul J, Robert Kurzban, Sheila Ahmadi, Ronald S Swerdloff, Jang Park, Levan Efremidze, Karen Redwine, Karla Morgan, and William Matzner (2009). "Testosterone administration decreases generosity in the ultimatum game". In: *PLOS ONE* 4.12, e8330. DOI: 10.1371/journal.pone.0008330 (cit. on pp. 85, 184).

Zethraeus, Niklas, Ljiljana Kocoska-Maras, Tore Ellingsen, BO Von Schoultz, Angelica Linden Hirschberg, and Magnus Johannesson (2009). "A randomized trial of the effect of estrogen and testosterone on economic behavior". In: *Proceedings of the National Academy of Sciences* 106.16, pp. 6535–6538. DOI: 10.1073/pnas.0812757106 (cit. on pp. 85, 182, 184).

Zhang, Jun and Kai F. Yu (1998). "What's the Relative Risk?: A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes". In: *JAMA* 280.19, p. 1690. DOI: 10.1001/jama.280.19.1690 (cit. on p. 78).

Zhang, Y Jane (2019). "Culture, Institutions and the Gender Gap in Competitive Inclination: Evidence from the Communist Experiment in China". In: *The Economic Journal* 129.617, pp. 509–552. DOI: 10.1111/ecoj.12596 (cit. on p. 84).

Zhong, Songfa, Idan Shalev, David Koh, Richard P. Ebstein, and Soo Hong Chew (2018). "Competitiveness and Stress". In: *International Economic Review* 59.3, pp. 1263–1281. DOI: 10.1111/iere.12303 (cit. on p. 84).

Ziel, Florian and Rick Steinert (2016). “Electricity price forecasting using sale and purchase curves: The X-Model”. In: *Energy Economics* 59, pp. 435–454. DOI: 10.1016/j.eneco.2016.08.008 (cit. on pp. 22, 53).

Ziel, Florian and Rick Steinert (2018). “Probabilistic mid- and long-term electricity price forecasting”. In: *Renewable and Sustainable Energy Reviews* 94, pp. 251–266. DOI: 10.1016/j.rser.2018.05.038 (cit. on pp. 22, 52).

Zweifel, Peter, Aaron Praktiknjo, and Georg Erdmann (2017). *Energy economics*. New York, NY: Springer Science+Business Media (cit. on pp. 23, 24).