Journal of Economic Behavior and Organization 188 (2021) 87–104

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Do non-choice data reveal economic preferences? Evidence from biometric data and compensation-scheme choice

Marja-Liisa Halko[a,b,*], Olli Lappalainen[c,d], Lauri Sääksvuori[e,f]

[a] Economics, P.O. Box 17, 00014 University of Helsinki, Finland
[b] Helsinki Graduate School of Economics, Finland
[c] Faculty of Management, University of Tampere, FI-33014 Tampere, Finland
[d] University of Turku, Department of Economics at Turku School of Economics, Rehtorinpellonkatu 3, 20500 Turku, Finland
[e] Finnish Institute for Health and Welfare, Centre for Health and Social Economics, P.O. Box 30, 00271 Helsinki, Finland
[f] University of Turku, INVEST Research Flagship Center, Assistentinkatu 7, 20014 University of Turku, Finland

ARTICLE INFO

Article history:
Received 21 January 2020
Revised 24 March 2021
Accepted 6 April 2021

JEL:
C91
D01
D03
J16
J24

Keywords:
Compensation schemes
Competition
Team
Experiment
Gender
Heart rate variability
Non-choice data

ABSTRACT

We investigate the feasibility of inferring economic choices from simple biometric non-choice data. We employ a machine learning approach to assess whether biometric data acquired during sleep, naturally occurring daily chores and participation in an experiment can reveal preferences for competitive and team-based compensation schemes. We find that biometric data acquired using wearable devices enable equally accurate out-of-sample prediction for compensation-scheme choice as gender and performance. Our results demonstrate the feasibility of inferring economic choices from simple biometric data without observing past decisions. However, we find that biometric data recorded in naturally occurring environments during daily chores and sleep add little value to out-of-sample predictions.

1. Introduction

Compensation schemes are expected to encourage commitment to increasing productivity and motivate employees to work toward organizational goals. Individual compensation plans generally include components related to productivity, relative performance evaluations and team payments, but the importance of these components may vary markedly between individuals, firms, and industries.

* Corresponding author.
E-mail address: marja-liisa.halko@helsinki.fi (M.-L. Halko).

https://doi.org/10.1016/j.jebo.2021.04.009
0167-2681/

Economic theory and empirical evidence suggests that relative compensation schemes have several desirable properties related to motivation, risk taking, and flexibility (Lazear and Rosen 1981; Nalebuff and Stiglitz 1983). Relative performance evaluations may also enable organizations to attract productive workers Lazear (2000). However, research has simultaneously suggested that competitive compensation schemes may discourage highly productive individuals (Niederle and Vesterlund 2007). Notably, men are often more likely to enter rank-order tournaments than women, even though there often are no gender differences in performance (Niederle and Vesterlund 2007, 2011).

We investigate self-selection into competitive and team-based incentive schemes. First, we relate several observable characteristics of individuals, including their past decisions, to their choice of compensation scheme. Second, we predict individuals' choice of compensation scheme using observable characteristics and biometric data on cardiac function. We test the feasibility of inferring choices from simple biometric data acquired using wearable technology. Moreover, we test an intriguing idea that biometric measures of bodily functions during sleep and daily chores serve as reliable predictors of economic choices.

Technological change has rapidly increased the availability of biometric data. This progress has expanded the set of potential data resources that can be used to predict decisions without observations of past choices. The addition of biometric data into the toolbox of economic analysis has not occurred without criticism (Gul and Pesendorfer 2008; Bernheim 2009). This critique commonly accepts that biometric data may help clarify how decisions are made but questions the added value of these non-choice data to identify causal relationships that show how economic choices are affected by different types of institutions. Moreover, biometric datasets, particularly neural data, have traditionally been cumbersome to acquire, and data collection often takes place in artificial environments, which may impede efforts to draw generalizable inferences for relevant policy questions and managerial decisions.

Similar to Harrison and List's (2004) terminology for economic experiments, our paper tests the feasibility of inferring choices from framed field data acquired using wearable technology during sleep and daily chores in naturally occurring environments. We evaluate the prospect of using biometric datasets collected in naturally occurring environments to predict real choices and test the out-of-sample predictive power of simple cardiac data on heart function compared to conventional economic choice data. We investigate whether simple cardiac data have predictive power similar to conventional revealed preference data and whether biometric data add predictive power to models that include such revealed preference data.

Our outcome data come from a laboratory experiment in which participants could choose among three compensation schemes: a piece-rate scheme, a competitive scheme, and a team-based compensation scheme. Our experimental design enabled us to observe participants' gender, performance, self-confidence and key economic preferences (risk and social preferences) that could all be related to their compensation-scheme choice. Thus, we were able to assess the relative predictive power of biometric data in contrast to several conventional economic variables that may be treated as observable or unobservable.

Our primary biometric measure is heart rate variability (HRV). HRV is as a well-established and clinically significant physiological phenomenon of variation in the time interval between heartbeats.[1] Simultaneously, HRV is often interpreted as a measure of autonomic nervous system activity that quantifies uncontrollable bodily responses to physiological and emotional stress. While voluminous empirical literature has associated short-term changes in HRV with acute emotional arousal, resting state HRV and long-term HRV measurements are often associated with more chronic emotional and physiological stress.[2] In applied research and occupational health care, long-term HRV measurements in naturally occurring situations are often used to assess physical and mental workload during a designated period of time, such as a work shift or specified portion of a work shift.[3] Overall, HRV has been empirically related to a wide range of physical health conditions and emotional states which makes it a relatively non-specific biometric marker for numerous potentially overlapping physiological and emotional phenomena.

Our experimental design that involves both short- and long-term biometric measurements is uniquely suited to test the predictive power of biometric non-choice data that can be acquired both during experimental conditions and naturally occurring daily chores. Being able to integrate simple, but predictive, measures of autonomic nervous system function in naturally occurring environments into the toolbox of economic analysis would open up new measurement techniques that are substantially more flexible than the prevailing neuroeconomic and biometric measurements conducted in laboratory environments.

[1] Resting state HRV is considered to be a robust predictor of mortality after cardiovascular disease events (Bigger et al. 1992). More broadly, a number of epidemiological studies tentatively suggest that HRV may serve as an independent predictor of future mortality and cardiovascular health in the general adult population (Huikuri et al., 1999, Dekker et al. 2000, Zulfigar et al., 2010). In the same vein, it has been hypothesized that low resting state HRV values could serve as a marker of less favorable health (Dekker et al., 2000). However, the predictive power and informative value of HRV recordings in generally healthy population samples and often used convenience samples, like undergraduate students, remains largely an open question.

[2] There is no clear and universal definition of emotional stress. There is also no clearly defined mapping between emotional stress and HRV measurements. However, there is voluminous empirical literature that has related HRV to various stress-related psychological measures (for reviews see e.g. Task Force 1996, Acharya et al. 2006, Järvelin-Pasanen et al. 2018).

[3] HRV recordings in naturally occurring environments have been used to assess the daily physiological and mental workload among several occupational groups including at least pilots (Roscoe 1992), truck drivers (Paxion et al. 2014), emergency physicians (Dutheil et al. 2012), surgeons (Dias et al. 2018) and firefighters (Kaikkonen et al. 2017). Fooken (2017) has compared HRV recordings during a university exam to HRV responses to common decision-making tasks in a laboratory experiment.

There are multiple reasons to investigate the predictive power of HRV data on compensation-scheme choice. First, HRV is often interpreted as an objective proxy for chronic and acute stress. Work-related stress is, in turn, claimed to associate with important career choices such as looking for a new job, declining a promotion or stepping down from managerial positions (APA 2007). Second, gender differences in self-selection into competitive incentive schemes is a major research theme which has produced systematic and robust evidence how key economic preferences relate to compensation-scheme choice. Thus, we are able to assess the relative predictive power of simple biometric data in contrast to several conventional economic variables in a research setting where these conventional variables are expected to be strong predictors of individual choices. Third, technological and cultural change that has led to a rising popularity of wearable health devices and biometric self-tracking readily generates vast amounts of data that may be of value to economic analysis and policy evaluation. Here we test the predictive power of these vast new data resources.

We investigate the predictive power of our non-choice data using machine learning methods. Several arguments motivate this machine learning approach. First, there are few theoretical constructs that researchers can harness to generate a priori testable predictions about potential relationships between physiological measures of stress and economic decisions.[4] Second, our biometric data create an exceptionally high-dimensional dataset with a large number of candidate variables that can serve as useful inputs for machine learning algorithms. Third, simple machine learning algorithms may substantially improve prediction accuracy and reduce overfitting by shrinking large regression coefficients and performing covariate selection. Thus, machine-learning methods can provide important tools that can enhance the utilization of non-choice data in predicting economic choices.

We find that men are substantially more likely to choose a competitive compensation scheme than women when given the option to enter a competitive, team-based, or piece-rate scheme. This result suggests that gender differences in competitiveness persist in situations in which options to compete and work in a team coexist. We also find that the choice of the team-based compensation scheme is largely guided by economic rationality. Our results show that low-performing men and women are more likely to choose a team than high-performing men and women.

More uniquely, our results suggest that simple biometric data collected using wearable devices during a decision process enable the prediction of economic choices. We find that biometric data acquired during the experimental paradigm enable greater predictive accuracy than a random classifier and generate for competitive compensation-scheme choice predictions that are as accurate as predictions based on gender and performance. However, our results show that biometric data recorded in naturally occurring environments during activities that are unrelated to the predicted outcome add little value to out-of-sample predictions. We find no evidence to support a conjecture that simple biometric measures collected during sleep predict economic choices.

Our results build on and contribute to several literatures. Our experimental design to elicit real choices over alternative compensation schemes builds on the experimental designs introduced by Niederle and Versterlund (2007) and Kuhn and Villeval (2015). We use simple cardiac data in conjunction with experimental data. Thus, our paper is a natural extension of the work that has previously explored the relationships between heart rate variability, competitiveness, and gender (Halko and Sääksvuori 2017). Overall, our empirical strategy can be viewed as an application of the non-choice revealed preference approach involving the estimation of statistical relationships between non-choice variables and real choices (Smith et al., 2014)

Our empirical results are largely consistent with the literature documenting the feasibility of inferring individual economic decisions from process data (Camerer 2007; Coricelli et al., 2010; Smith et al., 2014; Huseynov et al., 2019). However, our study provides a more comprehensive picture about the prospects and limits of using biometric data to predict economic behavior. We find that biometric data recorded in naturally occurring environments during activities that are unrelated to the predicted outcome add little value to out-of-sample prediction. Our results suggest that the feasibility of inferring economic decisions from simple cardiac data without observing past decisions is limited to measurements that are directly related to the predicted outcome.

Finally, our paper contributes to the discussion about the promises and perils of large datasets acquired using wearable technology. Our observation that researchers can predict individuals' economic choices with simple biometric data that is readily collected using mobile phones, wristbands, and smartwatches may offer new opportunities to use predictive analytics to forecast individual choices and to develop new preventative treatments to reduce undesired behaviors. Simultaneously, the growing body of evidence suggesting the existence of stable statistical relationships between simple biometric data and individual choices may help explain the demand for improvements in data security and highlights some potential risks related to the inconsiderate sharing of personal information.

The remainder of this paper is organized as follows: Section 2 presents our empirical strategy and data-collection methods, Section 3 presents our dataset, Section 4 summarizes our main empirical findings, and Section 5 provides our conclusions.

[4] A handful of studies have investigated the causal effects of stress on various economic behaviors, including time preferences (Haushofer et al. 2013, Koppel et al. 2017, Riis-Vestergard et al. 2018), risk preferences (Porcelli and Delgado 2009, Kandasamy et al. 2014, Cahlikova and Cingl 2017), and competitiveness (Buser et al. 2017, Zhong et al. 2018, Cahlikova et al. 2020, Esopo et al. 2019, Fu and Zhong 2019). These existing studies have produced largely mixed results and have not generated theoretical constructs that would enable to generate testable theory-based predictions how stress affects economic behavior in situations that have not yet been observed.

Table 1
Timing of experimental tasks.

	Task	Compensation scheme	Duration (min)
Task 0	Practice	–	2
Task 1	Adding numbers	Piece-rate	5
Task 2 / Task 3	Adding numbers	Competitive	5
Task 3 / Task 2	Adding numbers	Team-based	5
Task 4	Adding numbers	Piece-rate, competitive or team-based	5
Task 5	Choosing compensation scheme for Task 1	Piece-rate, competitive or team-based	~2
Belief elicitation	Rank guesses for Tasks 2–3	Payments from correct guesses	~2
Resting	Filling questionnaires	Payments from Holt & Laury	~10

Notes: The order of tasks 2 and 3 was counterbalanced across the experimental sessions.

2. Study design

This section documents our data-collection procedure and describes the variables used to examine our research questions. First, we describe the design of our behavioral experiment. Second, we describe the data-collection procedure for the cardiac data and the preprocessing of these data. Finally, we summarize key choice variables used in the results section.

2.1. Experimental design

Our experimental design consisted of two sessions over two consecutive days. We organized the first session to initiate the heart rate variability (HRV) measurements and elicit participants' social preferences. At the beginning of the first session, all participants agreed to wear an HRV recording device until the end of the second experimental session on the following day and to keep a record of their sleep patterns. We asked participants to attach an HRV measurement device to their chest (Electronic Supplementary Material, Section 1.1) and handed participants a take-home questionnaire (Electronic Supplementary Material, Section 3.1) to collect information about their sleep time, wake-up time, and potential interruptions in the HRV measurement. At the end of the first session, we elicited participants' social preferences using the social value orientation (SVO) measure (Electronic Supplementary Material Section 3.4, Murphy et al., 2011).

Our second experimental session and actual experimental design built on the experimental designs developed by Niederle and Versterlund (2007) and Kuhn and Villeval (2015). The basic task in our experiment was to add up sets of five two-digit numbers for five minutes. Participants solved these tasks individually. The experiment consisted of five arithmetic tasks under different compensation schemes (Table 1). Between the experimental tasks there was a two-minute long resting period during which the participants received feedback on their own performance in the pre-break task.

The compensation scheme in Task 1 was a non-competitive *piece-rate scheme*. Participants received a fixed payment of €0.25 for every correct answer. Participants did not receive any feedback about other participants' performance before completing all experimental tasks.

The compensation scheme in the second task (Task 2) was based on *competition* between four randomly assigned participants. The participant who correctly solved the most problems in a group was the winner and earned €1 per correct answer. All other participants earned nothing. In case of a tie, the winner was randomly chosen among the best performers.

The compensation scheme in the third task (Task 3) was based on performance of a *team*. Each participant was teamed with three other randomly assigned participants. Participants' earnings were based on the performance of all team members. Each team received a fixed payment of €0.25 for every correct answer by any team member. All accumulated earnings were divided equally among all members of a team. Thus, an individual i received a payment of $Y_i = (Q_1 + Q_2 + Q_3 + Q_4)/4$ for his or her work in Task 3. There were no efficiency advantages to team production. Before Task 2 and Task 3, we informed participants that the gender composition in the teams was mixed. The order of Tasks 2 and 3 was counterbalanced across the experimental sessions. In half of the sessions subjects did first Task 2 (competitive compensation), followed by Task 3 (team-based compensation), and in half of the sessions subjects did first Task 3, followed by Task 2.

In Task 4, participants selected which of the preceding three compensation schemes would apply to their future performance. If they chose the competitive scheme, their performance during Task 4 was compared against the performance of the other group members in Task 2. If they chose the team-based scheme, their performance during Task 4 was added to the output produced by other team members in Task 3. If they chose the piece-rate scheme, they received a fixed payment of €0.25 for every correct answer. This approach guaranteed that participants chose their preferred compensation scheme in Task 4 independent of other participants' compensation scheme choices.

Participants' choice of their preferred compensation scheme in Task 4 was the key outcome measure in our experiment. Our goal was to study participants' revealed preferences for compensation schemes in an environment where their choice set contained three different compensation schemes. Our approach with three relevant compensation schemes introduces a new important element to the literature on compensation-scheme choice.

We complemented our main outcome measure (Task 4) with a supporting outcome measure (Task 5). In Task 5, participants chose between the piece-rate, team-based, and competitive compensation schemes for their past performance in Task 1. Participants did not perform the arithmetic task but simply chose which compensation scheme would apply to their past

piece-rate performance. If they chose the competitive compensation scheme, their performance during Task 1 was compared with the Task 1 performance of the other Task 2 group members. If they chose the team-based compensation scheme, their performance in Task 1 was added to the Task 1 performance of the other Task 3 team members and earnings were divided equally among all members of the team. Task 5 had the same payoff structure as Task 4 but eliminated all aspects of performing in a team-based or competitive environment.

2.1.1. Confidence measures

Existing studies have documented that participants' confidence in their own ability and expectations about other participants' performance are crucial explanations of compensation-scheme choice. Consequently, our experiment included incentivized measures for participants' confidence and beliefs about other participants' performance under competitive and team-based compensations. We administered a short interim questionnaire asking participants to guess their rank in their group in Task 2 and to estimate the average number of correct solutions by other team members in Task 3. We rewarded participants if they were able to correctly estimate their rank in Task 2 and their team average (rounded to the nearest integer) in Task 3. The reward for a correct estimate was €1.

2.1.2. Risk elicitation

In theory, the optimal choice of compensation scheme depends on participants' attitude toward risk. Thus, we elicited participants' risk preferences using an incentivized measure for risk aversion (Holt and Laury 2002) and a general measure of risk-taking propensity derived from a one-item survey question (Dohmen et al., 2011). The full payoff table for the incentivized risk preferences measure is available in the Electronic Supplementary Material (Section 3.3).

2.2. Experimental procedure

We conducted the experimental sessions in the PCRC Experimental Laboratory at Turku University, Finland. There were 160 participants (72 male and 88 female). All participants participated in two experimental sessions as planned. There were 20 participants in each session. We arranged the first session at 4:00 p.m. and the second session at 10:00 a.m. on the following day. The vast majority of participants were young adults (average age: men 26.6 years, women 25.7 years) with no history of heart disease. All participants were non-smokers, and none took cardiovascular medication.[5]

At the beginning of the first session, we randomly assigned participants to visually isolated cubicles and asked them to attach the HRV measurement device to their skin. After all participants successfully attached the device, we elicited participants' social preferences using the SVO test (Murphy et al., 2011). The total payments from the first session included a €5 show-up fee and earnings from the incentivized SVO test (average payment €14.52, sd. = 3.36). The total duration of the first session was approximately 30 min.

At the beginning of the second session, we once again randomly assigned participants to visually isolated cubicles and delivered a hard copy of the experimental instructions.[6] We counterbalanced the order of Tasks 2 and 3 between the sessions. At the end of the second session, we paid participants based on one of the five tasks.[7] We showed participants five cards faced down on their computer screen and asked them to choose one. The cards were in a random order. The chosen card determined the relevant task for payment. The total payment in the second session included a €10 show-up fee, earnings from one randomly chosen task, earnings from the belief elicitation questions, and earnings from the incentivized measure for risk aversion (average payment: €18.38, sd. = 0.96). The total duration of the second session was approximately 60 min.

3. Dataset

This section describes some noteworthy characteristics of the data. In addition, the Electronic Supplementary Material documents our biometric data collection more extensively and details the construction of the various biometric variables (Electronic Supplementary Material, Section 1.2).

3.1. Choice data

Competitive and team-based compensation: The choice between competitive, team-based, and piece-rate compensation schemes generated our two key outcome measures. We created the variable *competition*, which took the value 1 if a participant chose the competitive compensation scheme and the value 0 if a participant chose either the team-based or piece-rate

[5] We recruited participants using the ORSEE software (Greiner, 2015). The experiment was programmed and conducted using the z-Tree software (Fischbacher, 2007). We informed potential participants in the invitation that the experiment involved non-invasive biometric monitoring and restricted participants to non-smokers and individuals who did not take cardiovascular medicine. The Ethics Committee of the Aalto University approved the study protocol and the study was conducted in accordance with the Helsinki Declaration. The analysis plan was not pre-registered.

[6] An English translation of the experimental instructions is available in the Electronic Supplementary Material (Section 3.2).

[7] Following the original study by Niederle and Vesterlund (2007), the economics literature on compensation-scheme choice largely implements an approach wherein one of the payoff-relevant tasks is randomly chosen for payment at the end of the experiment. Paying for only one task diminishes the chance that decisions are used to hedge against outcomes of other decisions. Charness et al. (2016) find that paying for only a subset of tasks is often as reliable as paying for all tasks.

compensation scheme. Similarly, we created the variable *team*, which took the value 1 if a participant chose the team-based compensation scheme and the value 0 if a participant chose either the competitive or piece-rate compensation scheme.

Overconfidence and beliefs: We created a measure for *overconfidence* using participants' answers to the incentivized interim questionnaire in which they guessed their rank in the tournament and data on their actual rank in the tournament. We computed the difference between the guessed rank and actual rank in the tournament and used this value as a measure for overconfidence.

Our incentivized interim questionnaire asked participants to estimate the average number of correct solutions by other team members under the team-based compensation scheme. Using these estimates and participants' own performance, we created the variable *better than the team average*, which took the value 1 if participants believed that their performance was better than the average performance in their own team and the value 0 if participants believed that their performance was worse than the average performance in their own team.

Risk attitude: The incentivized risk measurement lottery contained 10 paired lottery decisions with modest payoffs (Holt and Laury 2002). Participants had to make 10 successive choices between the two paired lotteries. The break-even point at which participants switched from the low-risk option to the high-risk option indicated their degree of risk aversion. In our general risk attitude question, participants were asked to indicate their general willingness to take risk on a scale from 0 to 10, where 0 stood for "not willing to take risk" and 10 stood for "completely willing to take risk".

Social preferences: The SVO test contained six resource-allocation decisions for which participants chose resource allocations between themselves and other anonymous participants (Murphy et al., 2011). Participants' allocations in the test indicated their SVO (altruistic, prosocial, individualistic, or competitive). We calculated the results from the SVO test and created the variable *prosocial*, which took the value 1 if a participant's SVO outcome was either prosocial or altruistic and the value 0 if a participant's SVO outcome was either individualistic or competitive.[8]

3.2. Biometric data

The HRV signal is a very high-dimensional biomedical signal with 1000 sampled values per second (1000 Hz). To obtain useful measures from the HRV signal, we reduced the very high-dimensional signal to a set of low-dimensional features. We used the Kubios HRV (v. 3.0.2) software to preprocess the HRV data and to construct our biometric variables (Tarvainen et al., 2014). To preprocess and correct potential artifacts in the data, we applied the software's automatic artifact-correction algorithm, which is based on observing deviations from the time-varying inter-beat-interval (IBI) value distribution (Tarvainen et al., 2019). Finally, we visually inspected the resulting corrected beat-to-beat interval series to detect potentially remaining outliers. At this preprocessing stage, we had to exclude seven participants from our biometric dataset either due to technical problems with the HRV measurement device during the measurement period or irreversible artifacts in the time series.[9]

We reduced the artifact-corrected IBI time series to a set of low-dimensional features using two analysis methods. The applied methods were based on the guidelines given by the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (Task Force, 1996). First, we used time-domain methods to derive the root mean square of the successive differences (RMSSD), which measures the variation of consecutive IBIs. Second, we used frequency-domain methods to decompose the IBI time series into a combination of different frequencies with varying power spectrum density. Our frequency-domain measure extracted from the power spectrum density was low frequency (LF: 0.04–0.15 Hz)/high frequency (HF: 0.15–0.4 Hz) power ratio (*LF/HF ratio*).

The wearable devices used to measure HRV also generate heart rate (HR) data. We supplemented our HRV measurements with HR data. HR is by definition negatively associated with HRV. This association is due to a physiological phenomenon and mechanistic relationship between the measures (Sacha 2014). The physiological relationship between HR and HRV is determined by the autonomic nervous system activity where higher parasympathetic nervous system activity leads to slower HR and higher HRV. HRV and HR are important, less than perfectly correlated, biometric measures of individuals' cardiac function and may serve as independent predictors of mortality and cardiovascular health (Huikuri et al., 1999).[10]

We not only reduced the HRV signal to a set of low-dimensional features but also divided the full time series into temporal segments (Fig. 1). The reduction of the HRV signal into low-dimensional features and the division of the time-series data into shorter segments created 39 biometric features that we used as input variables for our machine learning algorithms. A complete list of the biometric candidate variables and descriptive summary statistics are available in the Electronic Supplementary Material (Section 1.2.3).

[8] We used only one dummy variable because the SVO test results showed that 63.1% of the participants were classified as pro-social, 36.3% as individualistic, less than 1% as altruistic and no one as competitive. The share of participants classified as pro-social individuals (63.1%) falls within the range of previously reported results in the SVO literature. To compare our results to the previously reported share of pro-social individuals in the SVO literature, we extracted data from a meta-analysis summarizing the relationship between social value orientation and cooperation in social dilemmas (Pletzer et al., 2018). The meta-analysis includes 23 studies that use the SVO task with monetary incentives and report the percentage of subjects classified as pro-socials. Using these data, we find that the sample size weighted average share of pro-socials in these studies is 59.7 percent (95% CI: 54.7 − 64.8).

[9] We were not able to verify the definite cause of these irreversible artifacts in the HRV time series. However, the measurement devices we used, like many other wearable devices, are sensitive to potential problems with skin contact. In the event of inadequate skin contact, the HRV measurement device continues to record disturbed signals that cannot be corrected using statistical artifact-correction algorithms.

[10] There is a large literature that evaluates the predictive power of HR and HRV on mortality and cardiovascular health. However, the predictive power the combination of these two measures can provide compared to isolated measures of HRV and HR is largely an open question.

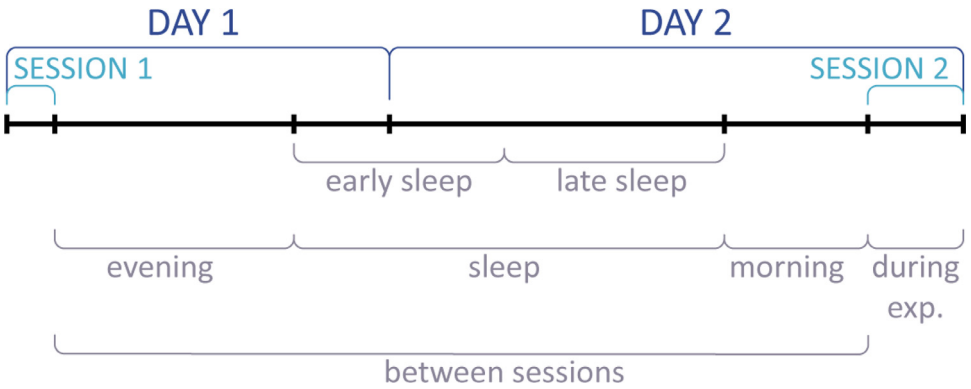

Fig. 1. Measurement periods of the biometric variables. The experiment started at 4 p.m. on day one and ended approximately at noon on day two. To construct our biometric variables we used six segments of the measurement period between the two sessions: the total time between the sessions, evening, the entire sleep time, the first and the second half of sleep time, and morning. The sleep length of an individual participant was based on the participant's report. We also measured changes in biometric variables during sleep, that is, the difference between late sleep and early sleep normalized by early sleep. The measurement period during the main experiment was divided into four segments: during the five minutes resting period before the first arithmetic task and during Tasks 1 to 3. We also measured participants' responses to competitive (team-based) compensation, that is, the difference between competitive (team-based) compensation and piece-rate compensation normalized by piece-rate compensation.

4. Results

This section summarizes our main empirical findings. First, we present descriptive statistics for the choice data. Second, we assess how individual characteristics, such as performance, overconfidence, risk attitude, and social preferences, relate to compensation-scheme choice using our performance measure and observable choice data from ancillary experimental tasks. Third, we use a machine learning approach to evaluate the feasibility of inferring choices from our biometric data. Finally, we report results from a machine learning based approach to find the most relevant biometric covariates associated with compensation-scheme choice.

4.1. Descriptive statistics

Table 2 presents descriptive statistics for the choice data. Under the piece-rate compensation scheme, the average number of correctly solved problems is 8.01. Under the tournament scheme, the average number of correctly solved problems is 9.37. Under the team-based compensation scheme, the average number of correctly solved problems is 8.59. We do not find any gender differences in average performance under different compensation schemes (Table 2: two-sided t-tests, $p > 0.155$).[11] We find that men and women performed significantly better under the tournament and team-based compensation schemes than under the piece-rate compensation scheme (Figure S2; Table 2: two-sided paired t-tests: tournament vs. piece rate for men $p = 0.002$, tournament vs. piece rate for women $p < 0.001$, team-based vs. piece rate for men $p = 0.060$, team-based vs. piece rate for women $p = 0.008$).

We find that participants were, on average, overconfident about their rank in the tournament. Their mean rank guess of 2.2 is significantly more optimistic than the true average rank of 2.5 (two-sided *t*-test against the true average rank of 2.5, $p = 0.001$). We find that men were more overconfident than women about their rank in the tournament ($p = 0.017$). However, we do not find that men were more overconfident than women about their performance in the team-based compensation scheme ($p = 0.151$). We document gender differences in risk attitudes. Men reported higher willingness to take risk in our general risk attitude questionnaire than women ($p = 0.026$) and chose a lower number of safe choices in an incentivized price list task designed to elicit risk aversion ($p = 0.097$).

Table 2 shows that the tournament compensation scheme was clearly the most preferred compensation scheme. We find that 44% of participants chose the tournament scheme, while 31% of participants chose the team-based scheme. The remaining 25% of participants preferred the piece-rate compensation scheme over the tournament and team-based schemes. Table 2 shows that 35% of women and 54% of men chose the tournament compensation scheme. Thus, there is a clear gender gap in tournament entry (Fisher's exact test, p-value $= 0.025$). We find that 34% of women and 26% of men chose the team-based compensation scheme (Fisher's exact test, p-value $= 0.307$). Thus, there is no significant gender gap in the choice of team-based compensation.

[11] In addition to not finding gender differences in average performance under different compensation schemes, we do not find gender differences in the probability of winning the tournament or in the probability of benefitting from the team-based compensation scheme. We assessed the probability of winning the tournament and benefitting from the team-based compensation through a simulation. The procedure for and results of the simulation are presented in detail in the Electronic Supplementary Material (Section 2).

Table 2
Descriptive statistics.

		Full sample		Women		Men		Women vs. men
Variable	Task/Stage	Mean	Sd.	Mean	Sd.	Mean	Sd.	P-value$^{\delta}$
Performance	Piece rate	8.01	3.19	7.68	2.76	8.40	3.62	.155
	Competition	9.37	3.09	9.38	3.03	9.36	3.19	.978
	Team	8.59	2.96	8.34	2.88	8.89	3.06	.246
	Own choice	10.43	3.41	10.28	3.30	10.60	3.57	.548
HRV (RMSSD)	Resting	36.14	19.63	38.09	22.14	33.73	15.84	.175
	Piece rate	29.48	14.39	30.80	16.15	27.88	11.84	.212
	Tournament	30.35	15.46	32.32	16.64	27.94	13.62	.081
	Team	32.17	16.55	34.28	19.45	29.59	11.76	.081
HR	Resting	80.86	11.52	80.39	10.59	81.44	12.64	.580
	Piece rate	85.34	11.77	84.59	11.34	86.25	12.29	.387
	Tournament	86.53	13.31	84.37	12.27	89.18	14.12	.026
	Team	82.50	11.38	81.45	10.56	83.79	12.27	.208
Rank guess		2.21	0.97	2.31	0.94	2.08	0.99	.146
Overconfidence		0.29	1.16	0.10	1.07	0.53	1.15	.017
Better than team		0.44	0.50	0.39	0.49	0.50	0.50	.151
Risk attitude (G)		4.63	2.33	4.26	2.28	5.08	2.33	.026
Risk attitude (H&L)		6.81	1.74	7.04	1.81	6.55	1.64	.097
Pro-sociality		0.64	0.48	0.63	0.49	0.65	0.48	.718
Competition		0.44	0.50	0.35	0.48	0.54	0.50	.025$^{\eta}$
Team		0.31	0.46	0.34	0.44	0.26	0.47	.307$^{\eta}$

Notes: This table summarizes the descriptive statistics for the choice variables and some key biometric variables in the full sample and for men and women separately. The reported statistics are averages and standard deviations. The first column displays the name of the variable. The last column reports p-values from a test that compares the average value between men and women. *Performance:* number of correctly solved problems during the task. *HRV (RMMSD):* participants' heart rate variability reported as the root mean square of successive differences (see Electronic Supplementary Material, Section 1.2.1, for the exact definition). *Resting* refers to the five-minute long resting period before the first arithmetic task. *HR:* participants' heart rate measured as beats per minute. *Rank guess*: participants' guessed rank in their group under the competitive compensation scheme. *Overconfidence*: the difference between guessed rank and actual rank in the tournament. *Better than team*: participant's belief that his/her performance in Task 3 was better than the average performance of the other team members = 1 and otherwise = 0. *Risk attitude (G)*: general risk attitude based on a survey question. *Risk attitude (H&L)*: risk attitude measured using the incentivized price list developed by Holt and Laury (2002). *Pro-sociality*: participants' scores on the SVO test. It is a dummy variable with prosocial or altruistic = 1 and individualistic or competitive = 0. δ = p-value for two-sided t-tests, η = p-value for Fisher's exact test.

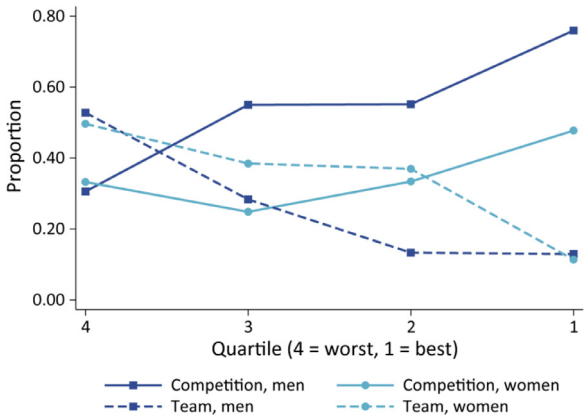

Fig. 2. The proportion of participants who selected the competitive compensation scheme (solid line) and the team-based compensation scheme (dashed line) conditional on their performance quartile under the piece-rate compensation scheme.

Fig. 2 shows the proportion of men and women who entered the competitive and team-based compensation schemes conditional on their performance quartile under the piece-rate compensation scheme. We find that there is a clear association between performance and the choice of compensation scheme. The proportion of men and women who entered the tournament increased with performance. Simultaneously, the proportion of men and women who entered the team-based compensation scheme decreased with performance. We observe that men were more likely to choose the competitive compensation scheme than women in all but the lowest performance quartile. At the same time, we observe no visible gender gap in willingness to enter the team-based compensation scheme in any performance quartile.

Table 3
Predictors of tournament entry - Logit models (1 = competitive, 0 = piece rate or team-based).

	(1)	(2)	(3)	(4)	(5)	(6)
Female	−0.199***	−0.202***	−0.124**	−0.213***	−0.199***	−0.125**
	(0.071)	(0.068)	(0.063)	(0.071)	(0.071)	(0.063)
Performance		0.047***	0.051***	0.046***	0.043***	0.039***
		(0.010)	(0.010)	(0.011)	(0.012)	(0.011)
Overconfidence			0.118***			0.099***
			(0.030)			(0.030)
Risk attitude			0.052***			0.055***
			(0.012)			(0.012)
Pro−sociality			−0.173***			−0.235***
			(0.059)			(0.063)
RMSSD, resting				0.002	0.001	0.002
				(0.002)	(0.002)	(0.001)
RMSSD change					−0.003**	−0.004***
(Tournament – Piece rate)					(0.002)	(0.001)
Observations	160	160	160	152	152	152
Pseudo R^2	0.043	0.111	0.294	0.118	0.141	0.343
Correctly classified (%)	63.75	66.25	73.75	64.47	64.47	75.66

Notes: This table reports the average marginal effects of logit models (standard errors in parentheses). *Female*: a dummy variable with female = 1 and male = 0. *Performance*: the number of correctly solved arithmetic problems under the competitive compensation scheme. *Overconfidence*: the difference between participants' true rank and rank guess. Values larger than 0 indicate overconfidence (true rank is worse than guessed rank). *Risk attitude*: participants' answers to a general risk attitude question on a scale from 0 to 10, where lower values indicate lower willingness to take risks. *Pro-sociality*: participants' scores on the SVO test. It is a dummy variable with prosocial or altruistic = 1 and individualistic or competitive = 0. *RMSSD resting* denotes RMSSD values during a five minutes resting period before the first arithmetic task. *RMSSD change* denotes competition-induced relative change in heart rate variability. All models include a control for the order of Tasks 2 and 3. Model 4 includes control variables for age and education ***Significant at $p < 0.01$, **Significant at $p < 0.05$, *Significant at $p < 0.1$.

Table 4
Predictors of tournament entry - Logit models for women and men separately (1 = competitive, 0 = piece rate or team-based).

	Women				Men			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Performance	0.045***	0.034**	0.045***	0.027	0.047***	0.056***	0.041***	0.044***
	(0.014)	(0.017)	(0.015)	(0.019)	(0.015)	(0.012)	(0.015)	(0.013)
Overconfidence		0.077*		0.069		0.149***		0.129***
		(0.047)		(0.046)		(0.038)		(0.045)
Risk attitude		0.057***		0.061***		0.062***		0.065***
		(0.016)		(0.017)		(0.020)		(0.018)
Pro−sociality		−0.274***		−0.314***		−0.020		−0.116
		(0.074)		(0.079)		(0.098)		(0.120)
RMSSD, resting			−0.0002	0.0006			0.005	0.004
			(0.002)	(0.002)			(0.003)	(0.003)
RMSSD change			−0.003	−0.003			−0.004*	−0.005**
(Tournament – Piece rate)			(0.002)	(0.002)			(0.002)	(0.002)
Observations	88	88	84	84	72	72	68	68
Pseudo R^2	0.066	0.270	0.093	0.319	0.123	0.335	0.172	0.404
Correctly classified (%)	61.36	78.41	63.10	80.95	66.67	80.56	67.65	79.41

Notes: This table reports the average marginal effects of logit models (standard errors in parentheses). *Performance*: the number of correctly solved arithmetic problems under the competitive compensation scheme. *Overconfidence*: the difference between participants' true rank and rank guess. Values larger than 0 indicate overconfidence (true rank is worse than guessed rank). *Risk attitude*: participants' answers to a general risk attitude question on a scale from 0 to 10, where lower values indicate lower willingness to take risks. *Pro-sociality*: participants' scores on the SVO test. It is a dummy variable with prosocial or altruistic = 1 and individualistic or competitive = 0. *RMSSD resting* denotes RMSSD values during a five minutes resting period before the first arithmetic task. *RMSSD change* denotes competition-induced relative change in heart rate variability. All models include a control for the order of Tasks 2 and 3. ***Significant at $p < 0.01$, **Significant at $p < 0.05$, *Significant at $p < 0.1$.

4.2. Compensation-scheme choice

In the following, we estimate how gender, performance, overconfidence, risk attitude, and social preferences relate to compensation-scheme choice. Table 3 reports the regression results.[12] Model 1 shows that the gender gap in tournament entry is 19.9 percentage points. Model 2 shows that women remained significantly less likely to enter the competitive

[12] There were three alternative compensation schemes as described in Section 2: a piece-rate scheme, a competitive scheme, and a team-based compensation scheme. Here our primary statistical model is one-vs-rest logistic regression. The coefficients and the odds ratios of the logit models are reported in the Electronic Supplementary Material (Table S6 and S7). In addition, results from an alternative multinomial logistic regression are reported in Table S9. We find that multinomial logit estimates are largely compatible with the logit estimates.

compensation scheme after controlling for performance. Model 3 shows that accounting for overconfidence, risk attitude, and social preferences reduce the gender gap from 20.2 percentage points to 12.4 percentage points. Overconfidence and risk attitude are positively and pro-sociality negatively associated with competitiveness. Taken together, these results suggest that overconfidence, risk attitude, and social preferences account for a substantial share of the gender gap in competitiveness when individuals are given the option to enter a tournament, team-based, or piece-rate scheme.[13]

Table 3 shows that men are substantially more likely to choose a competitive compensation-scheme than women in situations where there exists an option to enter a cooperative compensation-scheme besides a piece rate scheme. This result suggests that the regularly documented gender gap in competitiveness is robust to situations in which options to compete and cooperate coexist. In fact, we find that the size of the gender gap in competitiveness, 20 percentage points, matches the size of the gender gap that has been previously documented in the same pool of potential participants (Halko and Sääksvuori, 2017).[14] Moreover, the statistical relationships reported in Table 3 are largely consistent with the voluminous literature that has documented associations with various individual characteristics and competitive compensation-scheme choice.

More uniquely, our dataset enables an attempt to conceptually replicate previously documented associations between HRV and willingness to compete (Halko and Sääksvuori, 2017).[15] Using our present dataset, Table 3 (Models 4, 5 and 6) reports replications of previously reported associations between HRV and willingness to compete.[16] We find mixed evidence regarding the generalizability of previously reported results to the conditions of the present study. We find support for the result that competition-induced changes in heart rate variability are associated with participants' willingness to self-select into a competitive compensation-scheme (Table 3, Models 5 and 6). However, we are unable to find an association between baseline (resting state) HRV and willingness to compete in the context of this study (Table 3, Models 4, 5 and 6).[17]

Table 4 shows how performance, overconfidence, risk attitude, and social preferences relate to compensation-scheme choice by gender. Moreover, Table 4 reports replications of associations between HRV and willingness to compete separately for women and men. Table 4 (Models 3 and 4) show that resting state HRV is not a statistically significant predictor of tournament entry among women. Likewise, Table 4 (Models 7 and 8) show that resting state HRV is not a statistically significant predictor of tournament entry among men. The results for women in Table 4 do not replicate the previously reported result that women with high baseline HRV are more likely to choose tournament incentives than women with low baseline HRV.[18]

Table 4 (Model 4) shows that competition-induced change in heart rate variability is not associated with tournament entry among women. By contrast, Table 4 (Model 8) shows that competition-induced change in heart rate variability is a statistically significant predictor of tournament entry among men. Results in Table 4 are consistent with the results in Table 3 and replicate the previously reported result that men with large acute HRV response to competition are more likely to choose tournament incentives over piece rate incentives than men with small acute HRV response to competition.

We can only conjecture why evidence about the contextual replicability of previously reported results is mixed. We note that, even though the present experiment and the previously reported experiment by Halko and Sääksvuori (2017) have a comparable 5 min long resting-state measurement before participation in the experiment, there are numerous differences between experimental designs. Most importantly, participants in the present study were familiar with the experimental situation and had already been recording their HRV for several hours before participating in the second experimental session in which the relevant choice data were collected. These differences, among other factors, may explain the lack of replicability between the resting-state HRV and competitiveness in the context of the present study.

Table 5 reports regression results for willingness to enter the team-based compensation scheme. The results in Table 5 (Model 1) show that there are hardly any gender differences in entry into the team-based compensation payment

[13] The results are robust to controlling for observed socio-economic characteristics (age and education).

[14] The present paper and the previous study by Halko and Sääksvuori (2017) drew on the same pool of mainly student participants. We sent an invitation to participate in the present study only to subjects that had not participated in the previous study.

[15] Replication is often defined as an independent repetition the same scientific question using the original study's procedure to observe whether the previously reported findings recur. However, there are diverging views about the appropriateness of direct and conceptual replication attempts (see e.g. Zwaan et al. 2017, Nosek and Errington 2020). Due to numerous differences between experimental designs, our attempt to self-replicate our previously reported results is largely a by-product of the present study and can be seen as a conceptual replication attempt that enables us to study the generalizability of our previously reported findings (Halko & Sääksvuori, 2017).

[16] For consistency, all regression results reported in our paper are based on logit models. Following Halko and Sääksvuori (2017), Tables S10 and S11 in the Electronic Supplementary Material report associations between HRV and willingness to compete using probit models. Logit and probit models lead to similar qualitative conclusions. In addition, Table 3 (Models 3 and 6) reports an association between tournament entry and overconfidence (difference between the guessed rank and actual rank in the tournament), whereas Tables S10 and S11 report an association between tournament entry and participants' confidence in their performance, that is, their guessed rank. Including a variable that measures participants' confidence instead of a variable that measures overconfidence substantially reduces the gender gap in competitiveness (cf. Table 3 and S10), but has no effect on the strength of associations between HRV variables and willingness to compete

[17] We run a Monte-Carlo power analysis to understand how likely it is to observe in the current dataset previously reported effect sizes. The results from these simulations are summarized in Figure S7 and show that the current study with a sample size of 160 participants is adequately powered to detect previously reported associations between HRV data and compensations-scheme choice. Thus, we can with some confidence rule out an explanation that the lack of association between the baseline heart rate variability and willingness to compete in the present data is explained by inadequate power to detect the previously reported association.

[18] Halko and Sääksvuori (2017) report in their full dataset a point estimate of 0.006 (95% CI: 0.001 – 0.116) for the association between resting-state HRV and tournament entry and a point estimate of 0.009 (95% CI: 0.003 – 0.015) in a sample that contains only women.

Table 5
Predictors of team entry - Logit models (1 = team-based, 0 = piece rate or competitive).

	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.085	0.066	0.057	0.070	0.070	0.070
	(0.073)	(0.069)	(0.070)	(0.070)	(0.070)	(0.077)
Performance		−0.062***	−0.059***	−0.059***	−0.059***	−0.057***
		(0.012)	(0.014)	(0.012)	(0.012)	(0.016)
Better than team			0.031			0.022
			(0.080)			(0.087)
Risk attitude			−0.023*			−0.032*
			(0.014)			(0.016)
Pro−sociality			0.198***			0.306***
			(0.070)			(0.087)
RMSSD, resting				0.00007	0.00008	−0.0005
				(0.001)	(0.002)	(0.001)
RMSSD change					0.0001	0.0005
(Team – Piece rate)					(0.001)	(0.001)
Observations	160	160	160	152	152	152
Pseudo R^2	0.010	0.139	0.200	0.139	0.139	0.248
Correctly classified (%)	69.38	71.88	75.63	73.03	73.03	74.34

Notes: This table reports the average marginal effects of logit models (standard errors in parentheses). *Female*: a dummy variable with female = 1 and male = 0. *Performance*: the number of correctly solved arithmetic problems under the team compensation scheme. *Better than team*: participants believed their performance in Task 3 was better than the average performance of the other team members = 1 and otherwise = 0. *Risk attitude*: participants' answers to a general risk attitude question on a scale from 0 to 10, where lower values indicate lower willingness to take risks. *Pro-sociality*: participants' scores on the SVO test. It is a dummy variable with prosocial or altruistic = 1 and individualistic or competitive = 0. *RMSSD resting* denotes RMSSD values during a five minutes resting period before the first arithmetic task. *RMSSD change* denotes team-induced relative change in heart rate variability. All models include a control for the order of Tasks 2 and 3. Model 4 includes control variables for age and education. ***Significant at $p < 0.01$, **Significant at $p < 0.05$, *Significant at $p < 0.1$.

scheme. However, Table 5 (Model 2) shows that there is clear performance-based selection into the team-based compensation scheme. The higher their performance, the less likely participants were to choose the team-based compensation scheme. Table 5 (Model 3) shows that participants' pro-sociality is positively associated with entry into the team-based compensation scheme. Participants with prosocial preferences were 19.8 percentage points more likely to choose the team-based scheme than participants with non-prosocial preferences. Table 5 (Models 4, 5 and 6) shows that resting state HRV is not statistically significantly associated with team-based compensation-scheme choice. Likewise, Table 5 (Models 5 and 6) shows that individuals' HRV response to team-based incentives is not associated with team-based compensation-scheme choice. Taken together, these regression results suggest that variables that predict the choice of a competitive compensation scheme – gender, overconfidence, risk attitude, and HRV response to the change in compensation scheme – do not predict cooperativeness, which here seems to be associated with poor performance and prosocial preferences.[19]

4.3. Predictive accuracy

In Section 4.2, we found that gender, performance, overconfidence, risk attitude, and prosocial preferences relate to entry into a competitive compensation scheme and that performance and prosocial preferences are strongly related to entry into a team-based compensation scheme. We also found that there appears to be a statistical relationship between heart rate variability and compensation-scheme choice. However, all these associations were based on in-sample statistical fitting that does not provide information about the out-of-sample prediction performance of alternative statistical models based on various observable individual characteristics and biometric data recordings. Thus, in the following, we test how well alternative statistical models based on participants' gender, productive output, preferences, and biometric measurements predict compensation-scheme choice out of sample.

Our study generated a large amount of biometric data that may be used to predict entry into competitive and team-based compensation schemes. Our biometric measurements created a high-dimensional dataset that can be partitioned into a countless number of variables. We had no clear a priori theory about the relationship between unique biometric variables and compensation-scheme choice. Thus, we selected a large set of candidate variables recorded using the wearable sensors and tested the predictive value of these biometric candidate variables on compensation-scheme choice after controlling for overfitting.

[19] The results are robust to controlling for observed socio-economic characteristics (age and education). Table 4 and Table S8 in the Electronic Supplementary Material report logit models separately for women and men to explore potential gender differences in correlates of tournament and team-based compensation-scheme entry. Results from these estimations suggest that prosocial preferences are associated with competitive and team-based compensation-scheme choice only among women. Moreover, the association between overconfidence and tournament entry appears to be stronger among men than women.

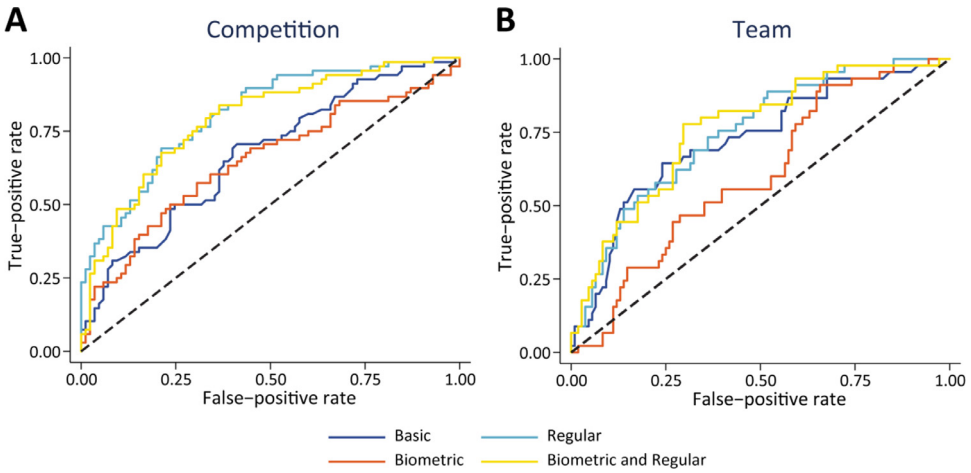

Fig. 3. ROC curves for predicting competitiveness (Panel A) and team-based compensation scheme choice (Panel B). The dashed lines (45-degree lines) represent the false- and true-positive rates of a random classifier. *The basic model* includes participants' gender and performance, *the regular model* includes participants' gender, performance, risk attitude, overconfidence, and social preferences, *the biometric model* includes all biometric candidate variables, and *the biometric & regular model* includes all biometric candidate variables and participants' gender, performance, risk attitude, overconfidence, and social preferences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We employed machine learning algorithms to test the predictive value of our biometric data on compensation-scheme choice. Our empirical strategy built on the research strategy introduced by Camerer et al. (2019).[20] We began with 39 biometric candidate variables. This set of candidate variables includes HRV measurements using time-domain methods, HRV measurements using frequency-domain methods and HR measurements during different stages of the measurement period. The units of measurement for these candidate variables (e.g. RMSSD, LH/HF-ratio) were selected based on the standard methods of HRV measurement defined by the Task Force (1996). The division of variables into waking hours and experimental hours was based on our research objective to investigate the predictive value of simple biometric data that is acquired during naturally occurring daily chores outside the laboratory environment. Our focus on measurements acquired during sleep was motivated by the well-known association between mental stress and sleep quality (e.g. Knudsen et al., 2007).[21] The full list of candidate variables is available in the Electronic Supplementary Material (Section 1.2.3).

The regression models in Tables 3 and show that gender and performance are strongly related to compensation-scheme choice. Simple prediction models including these two often observable characteristics create a natural reference model for model comparison. Simultaneously, we observe that risk aversion, overconfidence, and prosocial preferences relate to compensation-scheme choice. While these individual characteristics are often private and difficult to observe in naturally occurring environments, a more comprehensive prediction model that includes a broad set of conventional predictors for compensation-scheme choice creates an important reference point for models that rest on biometric data. Next we tested whether biometric data recorded using wearable devices would enable us to predict individuals' compensation-scheme choice in situations in which gender, performance, risk aversion, and confidence are considered unobservable. In addition, we tested whether biometric data would enable us to improve the performance of prediction models over the performance achieved using conventional economic measures related to compensation-scheme choice.

To assess the predictive power of biometric data, we created four alternative models to predict participants' competitiveness and team-based compensation scheme choice out of sample. First, our baseline model includes participants' gender and performance. Second, a more comprehensive model using conventional economic variables includes participants' gender, performance, risk attitude, overconfidence, and social preferences. Third, we estimated a prediction model using merely biometric data. Fourth, we estimated a prediction model that includes participants' gender, performance, risk attitude, overconfidence, social preferences, and biometric data.

We carried out our prediction models using the following nested cross-validation procedure (Figure S4). We divided our dataset into holdout and training samples. We separated two out of eight experimental sessions and used data from the remaining six sessions ($8 \times 7/2 = 28$ different splits in total) to train a linear model applying a least absolute shrinkage and

[20] A machine learning approach that selects a subset of relevant features from a large set of features using penalized regression methods has been widely used in bioinformatics, computer science, and neuroscience. Lately, there has been increasing interest in applying machine learning methods in economics (Krajbich et al. 2009, Belloni et al. 2012, Smith et al. 2014, Bajari et al. 2015, Peysakhovich and Naecker 2017). Varian (2014) and Mullainathan and Spiess (2017) provide general introductions to machine learning for economists. Camerer (2018) discusses how machine learning connects to behavioral economics.

[21] There are numerous potential approaches to quantify the association between mental stressors and sleep quality. Here we measure autonomous nervous system activity during sleep using HRV. Hall et al. (2004) show how experimentally induced stress affect HRV during sleep.

selection operator (LASSO) penalty.[22] The tuning parameter, λ, was optimized using five-fold cross-validation separately for every training set. Finally, we computed out-of-sample predictions for every test sample using the trained models.

We evaluated out-of-sample predictive accuracy using receiver operating characteristic (ROC) curves. ROC analysis is a standard tool in statistics used to quantify the performance of a binary classifier under different tradeoffs between false positives and false negatives. The use of ROC curves reflects the fact that one can always create more true positives (here, predictions for competitive and team-based compensation-scheme choices that do happen) but that doing so comes at the cost of predicting more false positives (here, predictions for entry into competitive and team-based compensation-schemes that do not happen). To compute the ROC curves, we compared our out-of-sample predictions for competitiveness and team-based compensation scheme at various decision threshold values, $t \in [0, 1]$, and plotted the true-positive rate against the false-positive rate at these threshold values. We classified all predicted values less than t as non-entry and all predicted values greater than or equal to t as entry. Thus, every point on the ROC curve represents the empirical false-positive and true-positive rates at the threshold.

ROC analysis enabled us to assess the performance of alternative classifiers over their entire operating range. For a random classifier, the true-positive and false-positive rates are identical, and the ROC curve is a 45-degree diagonal line. A well-performing classifier increases the true-positive rate (moving up on the y-axis) and decreases the false-positive rate (moving left on the x-axis). The most widely used measure for assessing prediction model performance is the area under curve (AUC), which measures the area under the ROC curve. We used AUC to compare the performance of our prediction models. The AUC for a random classier without predictive power is 0.5.

Fig. 3 shows ROC curves illustrating the relative trade-offs of predicting competitive and team-based compensation-scheme choices. For example, using our basic model (gender and performance) to classify competitiveness with 50 percent accuracy yields a false-positive rate (piece rate or team-based scheme incorrectly classified as competition) of 25 percent. We find that all models enable greater predictive accuracy than a random classifier. For competition (Panel A), a classifier that uses merely biometric data (AUC = 0.645, 95% CI: 0.554–0.735) is similar to a classifier that includes participants' gender and performance (AUC = 0.670, 95% CI: 0.584–0.756).[23] A classifier solely based on biometric data does not achieve the predictive accuracy of a classifier that is based on all standard predictors of competitiveness (AUC = 0.795, 95% CI: 0.725–0.865). The model that combines biometric and conventional data has slightly lower AUC (AUC = 0.777, 95% CI: 0.704–0.850) than the model with all standard predictors of competitiveness. Thus, adding biometric data to the conventional model of compensation-scheme choice does not decrease the out-of-sample prediction error and improve predictive accuracy.

For team-based compensation scheme (Panel B), a classifier that uses merely biometric data (AUC = 0.609, 95% CI: 0.513–0.705) enables greater predictive accuracy than a random classifier but does not achieve accuracy similar to a classifier that includes gender and performance (AUC = 0.734, 95% CI: 0.642–0.826). We observe that the ROC curve using biometric data is located substantially below the ROC curve based on all standard predictors of team-based compensation scheme (AUC = 0.768, 95% CI: 0.683–0.853). The model that combines biometric and conventional data (AUC = 0.760, 95% CI: 0.673–0.847) does not achieve significantly greater predictive accuracy than the model including all standard predictors of team-based compensation scheme.

Our results suggest that adding biometric data to the standard model does not increase the out-of-sample predictive accuracy for competitive and cooperative compensation-scheme choices. At the same time, we observe, as expected, that in-sample predictions are substantially more accurate for all models than the out-of-sample predictions (In-sample AUCs for predicting competitiveness: Basic model AUC = 0.711, 95% CI: 0.629–0.792; Regular model AUC = 0.826, 95% CI: 0.761–0.890; Biometric model AUC = 0.784, 95% CI: 0.709–0.858; Biometric & Regular model AUC= 0.878, 95% CI: 0.825–0.937). Moreover, we observe that the model that combines biometric and conventional data (Biometric & Regular) achieves substantially greater in-sample predictive accuracy than the model including all standard predictors of competitive compensation-scheme choice. Taken together, our observations suggest that adding biometric data to the standard model leads to overfitting and detection of statistical patterns that do not generalize to out-of-sample data.

The ROC analysis shows that biometric data enable greater out-of-sample predictive accuracy than a random classifier and generate for competitive compensation-scheme choice predictions that are as accurate as predictions based on gender and performance. In the following, we investigate the predictive value of our biometric measures more closely and divide the full measurement period into three stages: sleeping hours, waking hours, and experimental hours.[24] We investigate how the predictive accuracy of biometric data recorded during sleep (sleeping hours) and naturally occurring daily chores (waking hours) compare with the predictive accuracy of biometric data recorded during participation in the experiment (experimental hours).

[22] All independent variables were standardized (z-scored) to have a mean of 0 and a standard deviation of 1 prior to model training to account for the fact that the LASSO procedure is sensitive to the scale of inputs.

[23] We used AUC comparison tests (Delong et al., 1988) to formally evaluate the equality of out-of-sample predictive accuracy across all models. The results from these comparisons and related (Bonferroni-corrected) significance test statistics are reported in the Electronic Supplementary Material (Tables S12, S13 and S14). Our inferences concerning hypotheses that a predictive model performs no better than pure chance (AUC = 0.5) are more qualitative. Lieli and Hsu (2019) report that the AUC does not follow the usual asymptotic normal distribution and even bootstrap-based inference produces may lead to misleading results when testing the null hypothesis that AUC = 0.5.

[24] The division into sleeping and waking hours is based on individual surveys (Electronic Supplementary Material, Section 3.1). The reported sleeping hours accurately match expected changes in biometric measurements. For example, we observe that participants' HR is significantly lower during the reported sleeping hours (average HR: 60.9) than during the hours before sleep (average HR: 82.6) and after sleep (average HR: 88.3).

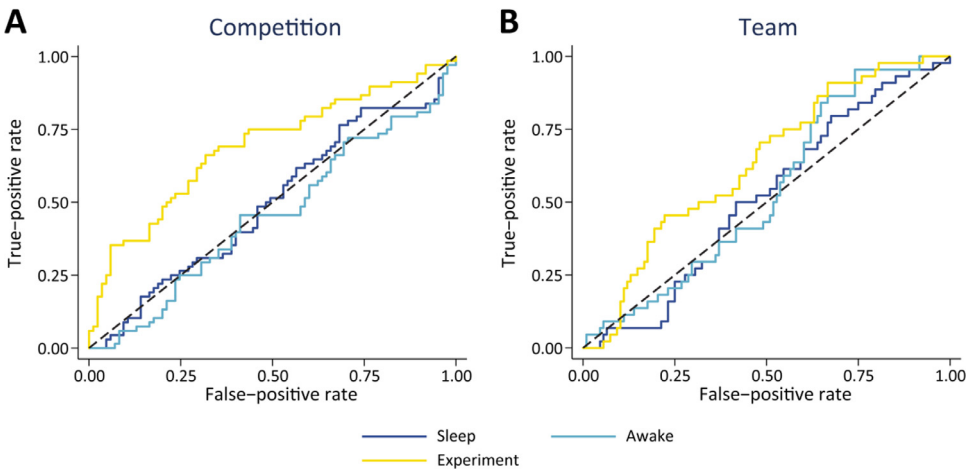

Fig. 4. ROC curves for predicting competitiveness (Panel A) and team-based compensation-scheme choice (Panel B) using solely biometric data. The dashed lines (45-degree lines) represent the false- and true-positive rates of a random classifier. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 4 shows separate ROC curves for classifiers that are solely based on biometric data measured during sleeping and waking hours. For competition (Panel A), we find that the classifiers excluding experimental hours trace the 45-degree line representing the false- and true-positive rates of a random classifier (sleeping hours: AUC = 0.504 95% CI: 0.410–0.597; waking hours: AUC = 0.437, 95% CI: 0.344–0.530). Similarly, for team-based compensation scheme (Panel B), the classifiers that include sleeping hours (AUC = 0.518, 95% CI: 0.420–0.615) and waking hours (AUC = 0.535, 95% CI: 0.440–0.630) trace the 45-degree line. These results suggest that biometric data recorded in naturally occurring environments during sleep and daily chores do not improve out-of-sample prediction.

Fig. 4 shows the ROC curves for a classifier based on participants' cardiac function during experimental tasks.[25] We observe that the ROC curves based on cardiac function during the experiment are substantially above the 45-degree line and the ROC curves for sleep and waking hours for competition (Panel A, AUC = 0.679, 95% CI: 0.592–0.766) and team-based compensation (Panel B, AUC = 0.638 95% CI: 0.545–0.731). These observations suggest that biometric signals for reliable out-of-sample prediction have to be recorded in environments that are sufficiently relevant for the predicted outcome.

4.4. Predictive variables

In Section 4.4, we found that classifiers using biometric data measured during active participation in the experiment outperform classifiers that use data recorded in naturally occurring environments. In the following, we complement these observations and investigate the best biometric predictors of compensation-scheme choice. The nested cross-validation procedure generated 28 model candidates by fitting a linear LASSO regression model to the 28 different training samples. To choose the best performing model, we calculated the ROC curves and corresponding AUC values for each model in the test sample using the estimates obtained from the related training sample, and selected the model with the highest AUC (e.g., see Sullivan Pepe 2003, Fawcett 2006). Finally, we fitted a logistic regression model to the entire dataset using the selected set of covariates with non-zero LASSO penalized coefficients (Belloni and Chernozhukov, 2013). We repeated the procedure twice – once for the choice of competitive compensation scheme and once for the choice of team-based compensation scheme.

Fig. 5 summarizes the marginal effects of logistic regression models that include all biometric covariates with non-zero LASSO penalized coefficients in the two best performing models. We find that participants' HR response to competition is *positively* associated with the choice of the competitive compensation scheme and *negatively* associated with the choice of the team-based compensation scheme. In addition, participants' HR in the evening (between the end of the first session and sleep) is *positively* associated with entry into the team-based compensation scheme.[26]

[25] Figures S5 and S6 in the Electronic Supplementary Material show ROC curves and report AUCs for classifiers based on biometric data separately for men and women to explore potential gender differences in the predictive power of biometric data. We do not detect any gender differences in the predictive power of biometric data for competitive compensation scheme choice. However, it turns out that classifiers based solely on biometric data predict team-based compensation-scheme choice somewhat better for men than for women (Fig. S6). A classifier that uses merely biometric data predicts women's team-based compensation-scheme choice no better than a random classifier.

[26] We carried out the search for best unique covariates using a model that included all biometric variables and participants' gender as a predictor. In the case of the team-based compensation scheme, the best performing model remains the same. In the case of the competitive compensation scheme, we find that participants' HR response to competition is still positively associated with the choice of the competitive compensation scheme. In addition, participants' HR and HRV responses to team-based compensation scheme are associated with competitive compensation-scheme choice in a model that contains all biometric variables and participants' gender (see Fig. S9).

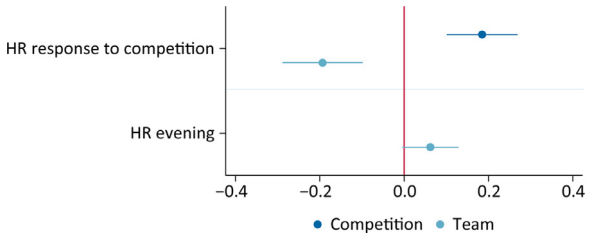

Fig. 5. Biometric variables selected using nested cross-validation and average marginal effects after fitting the logistic regression model that produced the highest AUC value. All variables are standardized to have a mean of 0 and a standard deviation of 1. Variable *HR response to competition* measures participant's heart rate response to competitive compensation and is measured as the difference between competitive compensation and piece-rate compensation normalized by piece-rate compensation. Variable *HR evening* measures participant's heart rate between the end of the first session and sleep. The figure presents the average marginal effects and 95% confidence intervals from two separate logistic regressions for outcome prediction. *The dark blue dots* represent the average marginal effects from a model where the dependent variable is the willingness to choose competition-based compensation scheme, and *the light blue dots* represent the average marginal effects from a model where the dependent variable is the willingness to choose team-based compensation scheme.

Fig. 5 shows that there are significant associations between compensation-scheme choice and HR response to the compensation scheme. However, the findings summarized in Fig. 5 are associations that do not allow causal interpretation. In addition, Fig. 5 may not include all covariates that are significantly associated with compensation-scheme choice. Our full set of candidate variables includes groupings of variables with high pairwise correlations, in which case LASSO tends to select only one variable from the group of highly correlated variables (Zou and Hastie, 2005).[27]

To overcome the limitations of LASSO, we reproduced the search for the best biometric predictors using Elastic net regularization that combines the penalty parameters of the LASSO and Ridge regularizations (Zou and Hastie, 2005). Even though the Elastic net regularization leads to a larger set of selected covariates in almost all 28 different splits of the data (Table S15) and to a larger set of selected covariates in a model with the highest AUC, we find that the same biometric covariates (HR response to competition and HR evening) that are selected using the LASSO regularization are selected using the Elastic net regularization (Figure S8).

Despite the limitations and differences of regularization methods, our search for the most predictive biometric variables indicates that participants' personal characteristics may relate to competitiveness and team-based compensation scheme choice. First, we observe that participants who chose to compete had a higher HR under the competitive compensation scheme than under the other compensation schemes (Table S4). In other words, there is a selection effect such that participants with a large competition-induced change in HR self-selected into the competitive compensation scheme (pairwise correlation $= 0.32$, $p < 0.000$). Second, participants who chose to enter the competitive compensation scheme had the largest difference in HR between the competitive and team-based compensation schemes (Table S4). Third, participants who chose team-based compensation scheme had higher HR in the evening (between the first experimental session and sleep) than other participants. Thus, there is selection effect such that participants with a high HR in the evening self-selected into the team-based compensation scheme (pairwise correlation $= 0.18$, $p = 0.025$).

5. Conclusions

This paper presents new evidence on the factors that predict self-selection into competitive and team-based compensation schemes. Beyond the literature on self-selection into competitive and team-based compensation schemes, our paper contributes to a nascent literature that investigates the feasibility of inferring economic choices from non-choice data. We extend the scope of inferring choices from non-choice data and acquire biometric data using wearable technology during sleep and daily chores in naturally occurring environments. Moreover, our biometric data are easier and less costly to obtain than the data used in many existing studies testing the predictive accuracy of non-choice data.

This paper replicates a regularly documented finding that men are more likely to choose a competitive compensation-scheme than women. Moreover, our result suggests that the gender gap in competitiveness is robust to situations in which options to compete and cooperate coexist. Our attempt to replicate previously documented associations between HRV and willingness to compete leads to more mixed results. We find support for a previously documented result that competition-induced changes in heart rate variability are associated with participants' willingness to self-select into a competitive compensation-scheme. However, we are unable to find an association between resting state HRV and willingness to compete using a previously defined 5-minute long resting-state HRV measurement. The mixed evidence from these replication efforts underlines the fact that there is no clear understanding or theory about the generalizability of these relationships and moti-

[27] Table S3 reports pairwise correlations between all biometric candidate variables and standard explanatory variables. We observe, as expected, that various time-domain HRV variables (RMSSD) are strongly correlated with each other and that various frequency-domain HRV variables (LFHF) are strongly correlated with each other. Likewise, various HR variables are strongly correlated with each other. Moreover, we observe, as explained in Section 3.2, that the time-domain HRV variables are strongly (negatively) associated with the HR variables. This strong grouping of variables is likely to explain the relatively small number of non-zero LASSO penalized coefficients in the model with the highest AUC value.

vates our machine learning approach that enables us to test the out-of-sample predictive accuracy of varied biometric data with little concern for overfitting.

Our results suggest that simple biological signals generate competitive compensation-scheme choice predictions that are as accurate as predictions based on gender and performance. However, we find that biometric data recorded in naturally occurring environments during sleep and daily chores do not improve out-of-sample prediction. More broadly, our results suggest that the feasibility of inferring economics decisions from simple non-choice data without observing past decisions may be limited to data that are directly related to predicted outcome.

Our machine learning approach shows that the measurement of physiological responses can lead to greater out-of-sample predictive accuracy than chance. Our results suggest that a larger number and different combinations of complementary physiological measures likely improve predictive accuracy beyond the accuracy reached using mere cardiac data and likely compensate for potentially unobservable choice data on past decisions. The use of machine learning approaches facilitates the search for predictive biometric variables and helps to avoid overfitting the prediction models.

Our results contribute to the debate on the promises and concerns of technological and cultural change stemming from the introduction of various biometric sensors (e.g., body temperature, skin conductivity, respiratory frequency, HR, and sleep quality) to mobile phones, wristbands, and smartwatches. We show that wearable data that can be linked to relevant economic activity has potential to contribute to an important objective of positive economics to forecast key economic choices and behaviors without interpreting the forecasting relationships as causal. Simultaneously, our results show that the bulk of cardiac data acquired using wearable technology during daily chores and sleep reveals little information about individuals' preferences, desires, and future choices.

We acknowledge that our study has several limitations. Our outcome variables were measured in a laboratory experiment. The artificiality of our decision environment may produce distorted choices that do not generalize to naturally-occurring decisions. While we acquired our biometric data using wearable technology not only during the decision-making process but also during sleep and daily chores in naturally-occurring environments, our participants were aware of the biometric data recording and may have responded in an unnatural manner to the fact that we were measuring their bodily functions. Our experimental instructions mentioned that the gender composition in the groups was mixed. Although the participants were also able to see that both men and women took part in the experiment, an explicit reference to gender composition could have served as a cue about appropriate behavior in the experiment and led to changes in behavior.

Our interest in this paper relates to testing the feasibility of inferring economic choices from biometric data without observing past decisions. Our study enables us to characterize factors that predict self-selection into competitive environments out-of-sample, but does not aim to improve our understanding on a question whether competitiveness exists as a separate trait independent of behavioral characteristics that predict self-selection into competitive incentive schemes (van Veldhuizen 2017; Gillen et al., 2019).

We note that the limitations related to the external validity of our study are not insuperable. It appears fairly straightforward to merge biometric data acquired using wearable technology to choice data (e.g., retail scanner data on purchases) that readily accumulates in naturally occurring situations. Likewise, there are already technological solutions to remotely measure vital bodily functions in the absence of potential experimenter demand effects. A natural next step toward understanding the full potential and relevancy of non-choice data for economics is to bring the data collection and outcome measurement from the laboratory to the field.

Acknowledgements

The authors want to thank Jonas Fooken, Tuukka Holster, Peter Matthews, and Henri Nyberg for useful discussion and comments and Juho Halonen for research assistance. This research is supported by the research project aivoAALTO (Aalto University), the Yrjö Jahnsson Foundation, and the Research Foundation of Cooperative Banks, Finland.

The authors declare no conflict of interest.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2021.04.009.

References

Acharya, U.R., Joseph, K.P., Kannathal, N., Lim, C.M., Suri, J.S., 2006. Heart rate variability: a review. Med. Biol. Eng. Comput. 44 (12), 1031–1051.
APA - American Psychological Association, 2007. Stress in America. American Psychological Association, Washington, DC.
Bajari, P., Nekipelov, D., Ryan, S.P., Yang, M., 2015. Machine learning methods for demand estimation. Am. Econ. Rev. 105 (5), 481–485.
Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80 (6), 2369–2429.
Belloni, A., Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. Bernoulli 19 (2), 521–547.
Bernheim, B.D., 2009. On the potential of neuroeconomics: a critical (but hopeful) appraisal. Ame. Econ. J. 1 (2), 1–41.
Bigger Jr, J.T., Fleiss, J.L., Steinman, R.C., Rolnitzky, L.M., Kleiger, R.E., Rottman, J.N, 1992. Frequency domain measures of heart period variability and mortality after myocardial infarction. Circulation 85 (1), 164–171.
Buser, T., Dreber, A., Mollerstrom, J., 2017. The impact of stress on tournament entry. Exper. Econ. 20 (2), 506–530.
Cahlíková, J., Cingl, L., 2017. Risk preferences under acute stress. Exper. Econ. 20 (1), 209–236.
Cahlíková, J., Cingl, L., Levely, I., 2020. How stress affects performance and competitiveness across gender. Manage Sci. 66 (8), 3295–3798.

Camerer, C.F., 2007. Neuroeconomics: using neuroscience to make economic predictions. Econ. J. 117 (519), C26–C42.
Camerer, C.F., 2018. Artificial intelligence and behavioral economics. The Economics of Artificial Intelligence: An Agenda. University of Chicago Press.
Camerer, C.F., Nave, G., Smith, A., 2019. Dynamic unstructured bargaining with private information: theory, experiment, and outcome prediction via machine learning. Manage Sci 65 (4), 1455–1947.
Charness, G., Gneezy, U., Halladay, B., 2016. Experimental methods: pay one or pay all. J. Econ. Behav. Organ. 131, 141–150.
Coricelli, G., Joffily, M., Montmarquette, C., Villeval, M.C., 2010. Cheating, emotions, and rationality: an experiment on tax evasion. Exper. Econ. 13 (2), 226–247.
DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves. Biometrics 44 (3), 837–845.
Dekker, J.M., Crow, R.S., Folsom, A.R., Hannan, P.J., Liao, D., Swenne, C.A., Schouten, E.G., 2000. Low heart rate variability in a 2-minute rhythm strip predicts risk of coronary heart disease and mortality from several causes: the ARIC Study. Circulation 102 (11), 1239–1244.
Dias, R.D., Ngo-Howard, M.C., Boskovski, M.T., Zenati, M.A., Yule, S.J., 2018. Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. Br. J. Surg. 105 (5), 491–501.
Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., Wagner, G.G., 2011. Individual risk attitudes: measurement, determinants, and behavioral consequences. J. Eur. Econ. Assoc. 9 (3), 522–550.
Dutheil, F., Boudet, G., Perrier, C., Lac, G., Ouchchane, L., Chamoux, A., Schmidt, J., 2012. JOBSTRESS study: comparison of heart rate variability in emergency physicians working a 24-hour shift or a 14-hour night shift—A randomized trial. Int. J. Cardiol. 158 (2), 322–325.
Esopo, K., Haushofer, J., Kleppin, L., Skarpeid, I., 2019. Acute Stress Decreases Competitiveness Among Men *Working paper*.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874. doi:10.1016/j.patrec.2005.10.010.
Fischbacher, U., 2007. Z-tree, Zurich toolbox for ready-made economic experiments. Exper. Econ. 10, 171–178.
Fooken, J., 2017. Heart rate variability indicates emotional value during pro-social economic laboratory decisions with large external validity. Sci. Rep. 7, 44471.
Fu, J., Zhong, S., 2019. Visceral Influences and Gender Difference in Competitiveness Available at SSRN 3341678.
Gillen, B., Snowberg, E., Yariv, L., 2019. Experimenting with measurement error: techniques with applications to the Caltech cohort study. J. Polit. Econ. 127 (4), 1826–1863.
Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. J. Econ. Sci. Ass. 1, 114–125.
Gul, F., Pesendorfer, W., 2008. The case for mindless economics. Found. Positive Normative Econ. 1, 3–42.
Halko, M.-L., Sääksvuori, L., 2017. Competitive behavior, stress, and gender. J. Econ. Behav. Org. 141, 96–109. doi:10.1016/j.jebo.2017.06.014.
Hall, M., Vasko, R., Buysse, D., Ombao, H., Chen, Q., Cashmere, J.D., Thayer, J.F., 2004. Acute stress affects heart rate variability during sleep. Psychosom. Med. 66 (1), 56–62.
Harrison, G.W., List, J.A., 2004. Field experiments. J. Econ. Lit 42 (4), 1009–1055.
Haushofer, J., Cornelisse, S., Seinstra, M., Fehr, E., Joëls, M., Kalenscher, T., 2013. No effects of psychosocial stress on intertemporal choice. PLoS One 8 (11), e78597.
Holt, C.A., Laury, S.K., 2002. Risk aversion and incentive effects. Am. Econ. Rev. 92 (5), 1644–1655.
Huikuri, H.V., Mäkikallio, T., Airaksinen, K.J., Mitrani, R., Castellanos, A., Myerburg, R.J., 1999. Measurement of heart rate variability: a clinical tool or a research toy? J. Am. Coll. Cardiol. 34 (7), 1878–1883.
Huseynov, S., Kassas, B., Segovia, M.S., Palma, M.A., 2019. Incorporating biometric data in models of consumer choice. Appl. Econ. 51 (14), 1514–1531.
Järvelin-Pasanen, S., Sinikallio, S., Tarvainen, M.P., 2018. Heart rate variability and occupational stress - systematic review. Ind Health 56 (6), 500–511.
Kaikkonen, P., Lindholm, H., Lusa, S., 2017. Physiological load and psychological stress during a 24-hour work shift among Finnish firefighters. J. Occup. Environ. Med. 59 (1), 41–46.
Kandasamy, N., Hardy, B., Page, L., Schaffner, M., Graggaber, J., Powlson, A.S., Coates, J., 2014. Cortisol shifts financial risk preferences. Proc. Natl. Acad. Sci. 111 (9), 3608–3613.
Knudsen, H.K., Ducharme, L.J., Roman, P.M., 2007. Job stress and poor sleep quality: data from an American sample of full-time workers. Soc. Sci. Med. 64 (10), 1997–2007.
Koppel, L., Andersson, D., Posadzy, K., Västfjäll, D., Tinghög, G., 2017. The effect of acute pain on risky and intertemporal choice. Exper. Econ. 20 (4), 878–893.
Krajbich, I., Camerer, C., Ledyard, J., Rangel, A., 2009. Using neural measures of economic value to solve the public goods free-rider problem. Science 326 (5952), 596–599.
Kuhn, P., Villeval, M.C., 2015. Are women more attracted to co-operation than men? Econ. J. 125 (582), 115–140. doi:10.1111/ecoj.12122.
Lazear, E.P., 2000. Performance pay and productivity. Am. Econ. Rev. 90 (5), 1346–1361.
Lazear, E.P., Rosen, S., 1981. Rank-order tournaments as optimum labor contracts. J. Polit. Econ. 89 (5), 841–864.
Lieli, R.P., Hsu, Y.C., 2019. Using the area under an estimated ROC curve to test the adequacy of binary predictors. J. Nonparametr Stat. 31 (1), 100–130.
Mullainathan, S., Spiess, J., 2017. Machine learning: an applied econometric approach. J. Econ. Perspect. 31 (2), 87–106.
Murphy, R.O., Ackermann, K.A., Handgraaf, M., 2011. Measuring social value orientation. Judgm Decis Mak 6 (8), 771–781.
Nalebuff, B.J., Stiglitz, J.E., 1983. Prizes and incentivess: towards a general theory of compensation and competition. The Bell J. Econ. 21–43.
Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? Do men compete too much? Q. J. Econ. 122 (3), 1067–1101.
Niederle, M., Vesterlund, L., 2011. Gender and competition. Ann. Rev. Econom. 3 (1), 601–630.
Nosek, B.A., Errington, T.M., 2020. What is replication? PLoS Biol. 18 (3), e3000691.
Paxion, J., Galy, E., Berthelon, C., 2014. Mental workload and driving. Front. Psychol. 5, 1344. doi:10.3389/fpsyg.2014.01344.
Peysakhovich, A., Naecker, J., 2017. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. J. Econ. Behav. Organ. 133, 373–384.
Pletzer, J.L., Balliet, D., Joireman, J., Kuhlman, D.M., Voelpel, S.C., van Lange, P.A.M., 2018. Social value orientation, expectations, and cooperation in social dilemmas: a meta-analysis. Eur. J. Pers. 32, 62–83.
Porcelli, A.J., Delgado, M.R., 2009. Acute stress modulates risk taking in financial decision making. Psychol. Sci. 20 (3), 278–283.
Riis-Vestergaard, M.I., van Ast, V., Cornelisse, S., Joëls, M., Haushofer, J., 2018. The effect of hydrocortisone administration on intertemporal choice. Psychoneuroendocrinology 88, 173–182.
Roscoe, A.H., 1992. Assessing pilot workload. Why measure heart rate, HRV and respiration? Biol. Psychol. 34 (2–3), 259–287.
Sacha, J., 2014. Interaction between heart rate and heart rate variability. Ann. Noninvasive Electrocardiol. 19 (3), 207–216.
Smith, A., Bernheim, B.D., Camerer, C.F., Rangel, A., 2014. Neural activity reveals preferences without choices. American Econ. J. 6 (2), 1–36.
Sullivan Pepe, M., 2003. The Statistical Evaluation of Medical Tests For Classification and Prediction. Oxford University Press, Oxford.
Tarvainen, M.P., Niskanen, J.-P., Lipponen, J.A., Ranta-aho, P.O., Karjalainen, P.A., 2014. Kubios HRV – heart rate variability analysis software. Comput. Methods Programs Biomed. 113 (1), 210–220.
Tarvainen, M.P., Lipponen, J., Niskanen, J.-P., Ranta-aho, P.O., 2019. Kubios HRV User's Guide https://www.kubios.com/downloads/Kubios_HRV_Users_Guide.pdf.
Camm, A.J., Malik, M., Bigger, J.T., Breithardt, G., Cerutti, S., Cohen, R.J., Lombardi, F.Task Force of the European Society of Cardiology the North American Society of Pacing, 1996. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Circulation 93 (5), 1043–1065.
van Veldhuizen, R., 2017. Gender Differences in Tournament Choices: Risk preferences, Overconfidence or Competitiveness?, *Discussion Paper, No. 14*. Ludwig Maximilians-Universität München und Humboldt-Universität zu Berlin, Collaborative Research Center Transregio 190 - Rationality and Competition, München und Berlin.
Varian, H.R., 2014. Big data: new tricks for econometrics. J. Econ. Perspect. 28 (2), 3–28.

Zhong, S., Shalev, I., Koh, D., Ebstein, R.P., Chew, S.H., 2018. Competitiveness and stress. Int. Econ. Rev. (Philadelphia) 59 (3), 1263–1281.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Royal Stat. Soc. 67 (2), 301–320.
Zulfiqar, U., Jurivich, D.A., Gao, W., Singer, D.H., 2010. Relation of high heart rate variability to healthy longevity. Am. J. Cardiol. 105 (8), 1181–1185.
Zwaan, R., Etz, A., Lucas, R., Donnellan, B., 2017. Making replication mainstream. Behav. Brain Sci. 1–50.