Butschek, Sebastian; González Amor, Roberto; Kampkötter, Patrick; Sliwka, Dirk

Working Paper

Paying Gig Workers – Evidence from a Field Experiment

IZA Discussion Papers, No. 12667

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Butschek, Sebastian; González Amor, Roberto; Kampkötter, Patrick; Sliwka, Dirk (2019) : Paying Gig Workers – Evidence from a Field Experiment, IZA Discussion Papers, No. 12667, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/207491

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12667

Paying Gig Workers – Evidence from a Field Experiment

Sebastian Butschek
Roberto González Amor
Patrick Kampkötter
Dirk Sliwka

OCTOBER 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12667

Paying Gig Workers – Evidence from a Field Experiment

Sebastian Butschek
University of Cologne

Patrick Kampkötter
University of Tübingen

Roberto González Amor
Zalon by Zalando

Dirk Sliwka
University of Cologne, CESifo and IZA

OCTOBER 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Paying Gig Workers – Evidence from a Field Experiment*

We study the performance effects of payment schemes for freelancers offering services on an online platform in an RCT. Under the initial scheme, the firm pays workers a pure sales commission. The intervention reduces the commission rate and adds a fixed payment per processed order to insure workers against earnings risk. Our experiment tests predictions from a formal model on labor supply and performance for individuals with different degrees of risk aversion and intrinsic motivation for the task. The treatment did not affect labor supply and even though the commission rate was reduced by 50% we find no sizeable loss in sales per order. However, there is strong evidence for heterogeneous treatment effects. The treatment reduced performance for less intrinsically motivated workers. For more intrinsically motivated workers, however, we observe the opposite pattern as performance increased even though commission rates were reduced.

JEL Classification: D23, J33, M52

Keywords: incentives, risk aversion, intrinsic motivation, sales compensation, multitasking, field experiment, gig economy, on demand economy, platform economy

Corresponding author:
Dirk Sliwka
University of Cologne
Faculty of Management, Economics and Social Sciences
Albertus-Magnus-Platz
50923 Köln
Germany

E-mail: dirk.sliwka@uni-koeln.de

* We thank Matthias Heinz and Timo Vogelsang as well as seminar and conference participants at the IZA/OECD Workshop "Labor Productivity and the Digital Economy" in Paris, the workshop "Natural Experiments and Controlled Field Studies" in Ohlstadt, Colloquium on Personnel Economics in Zurich, ESPE in Glasgow, EEA ESEM in Lisbon and at seminars in Cologne and Mainz for valuable comments and suggestions. We also thank the representatives of Zalando and Zalon, particularly Ivo, Rene and Diana, for their support. Helpful research assistance was provided by Savina Häfele, Sarah Helene Schaar, Sophia Schneider, and Stefanie Schweitzer. We have obtained an IRB approval from the University of Tuebingen (A2.5.4_065_aa). This study is registered with the AEA RCT Registry under registration number AEARCTR-0001555. We thank the German Research Foundation (DFG) for financial support through research unit "Design and Behavior – Economic Engineering of Firms and Markets" (FOR 1371).

1 Introduction

The importance of freelance work has been growing. In a survey for the US Katz and Krueger (2019) find that the fraction of workers in alternative work arrangements rose from 10.7 percent in 2005 to 15.8 percent in 2015. Part of this increase can be ascribed to the rise of the gig economy: a growing share of freelance suppliers of work are matched to customers on online platforms (Kuhn and Maleki (2017)). Tracking the number of open projects and tasks posted on a sample of such platforms, Kässi and Lehdonvirta (2018) document that the demand for online gig work has grown by about 21% from May 2016 to January 2018.[1]

One of the key differences between classical employment relations and freelance work is the flexibility of labor supply. While in classical employment relations are bound by contractual working hours over considerable time horizons, freelancers can frequently vary the quantity of work they offer or the number of assignments they accept. Firms employing them therefore face a dual incentive problem: ensuring both the quality of work in each assignment and getting freelancers to provide enough labor.[2] This results in a trade-off related to the classical trade-off between risk and incentives. When objective performance measures are available, freelancer effort on a given assignment is best incentivized by paying them a pure commission. Yet, such a payment scheme induces income uncertainty for freelancers that may reduce their incentives to accept a sufficient number of assignments. In turn, it may be optimal to reduce commission rates and introduce a fixed payment per assignment. We argue that this trade-off between incentives for performance and labor supply may not only be affected by freelancers' risk aversion, but also by their intrinsic motivation for the task at hand: Workers with a high task motivation should reduce their efforts to a weaker extent when the commission rate is reduced.

To test our claims we implement a natural field experiment in the gig economy. Our intervention changes freelancer compensation in a way that insures them against earnings fluctuations. This allows us to address two research questions. The first focuses on our setting, where the firm's objective is to increase freelancers' labor supply without substantial performance losses: given workers' preferences, should the firm continue to pay workers a pure commission or reduce the commission and add a fixed, certain payment per order? The second question is broader and has policy implications in other settings where freelance or atypical work is also common: how do workers with different task motivations and personality traits respond to compensation that reduces their earnings uncertainty but also their material performance incentives?

We first present a principal-agent model that captures a central feature of both our empirical setting and freelance work more generally: workers' ability to choose both how much work to provide and how much effort to invest. Our pre-registered hypotheses predict that the intervention will increase stylists' labor supply and decrease their sales performance overall. They further predict that greater risk aversion will be associated with larger labor supply increases while the negative sales performance effect will be less pronounced the greater a worker's intrinsic motivation.

For the field experiment we partnered with Zalon, an online provider of curated fashion shopping. The firm

[1] Abraham et al. (2017) provide a typology of work arrangements and define gig work (or on demand work or platform work) as a job meeting none of the following criteria: paid wage or salary, implicit or explicit contract for continuing relationship, predictable work schedule and predictable earnings when working.

[2] The second problem could be viewed as less severe as platforms may spread the work among many people (each providing little labor). However, adding workers to the platform will nearly always incur fixed costs of employee selection. Hence, there is an incentive to also raise the labor supply per worker.

is part of Zalando, the largest online fashion retailer in Europe.[3] Zalon relies on freelancers to provide remote styling services to fashion shoppers. It pays these stylists a pure commission on realized sales. Unlike gig workers on other platforms, Zalon's freelance fashion consultants are selected based on formal qualifications for the styling service and on their fashion affinity. Many of them, for example, are designers, tailors, and fashion students. As a consequence, intrinsic motivation is likely to play an important role in their freelance work. This differentiates our study from previous papers on the gig economy, which have focused on low-skilled, standardized and repetitive tasks.

All 202 new stylists starting work for Zalon between October 2016 and August 2017 participated in our RCT without knowing that they were part of a field experiment. We randomly allocated new stylists to the treatment or the control group sequentially - as they were hired. To increase power, we did so within four predicted labor supply strata. The individual treatment period consisted of the first two calendar months of stylists' work for Zalon. Control group stylists received a pure commission of 15% of the sales per order they generated. Treatment group stylists received a commission of 7.5% of sales per order plus a piece rate of €6.50 per order. We elicited workers' risk aversion and intrinsic motivation for the task in a pre-experimental online questionnaire. Labor supply and sales performance are measured using Zalon's order-level administrative data. Our proxy for labor supply is the total number of desired slots of each stylist during the treatment period; our measure for sales performance is net merchandise value (NMV) per order in Euro. Our setting is particularly suited for a field experiment as spillover effects are substantially less likely to appear than in typical firm-level field experiments, as (i) stylists work independently from home and (ii) we only include newly hired stylists which hardly get in contact with each other.

We find that overall, the intervention did not significantly increase average labor supply. Moreover, even though the commission rate was reduced by 50% and thus marginal incentives per "gig" were substantially lower, we find no sizeable loss in sales per gig. While the labor supply response did not vary significantly by risk aversion, we find strong support for our hypothesis of a heterogeneous treatment effect on sales performance by intrinsic information and economically sizeable differences. More specifically, and in line with standard reasoning, among less intrinsically motivated stylists (as classified by a median split), the intervention reduced performance by 17%. For the intrinsically more motivated stylists, however, the intervention even increased performance by 10%. Hence, in this group, the intervention not only mitigated the loss in material incentives as we hypothesized, but it even led to a higher motivation to perform. This is not only true in our main specifications, but also when we take into account that intrinsic motivation may be confounded by other personality traits or lower ability. It also holds when we consider customers' repurchase rate as an alternative performance measure: In the group of workers with above-median intrinsic motivation for the task as measured by our pre-experimental survey, the treatment raised repurchase rates by 41%, while the treatment reduced repurchase rates by 53% in the less motivated group.

We explore why our intervention did not increase labor supply. The most likely reason is that – while the treatment led to a better insurance against income uncertainty – it reduced workers' expected earnings per customer. Comparing the level of risk aversion of the workers in Zalon with data from two representative surveys of the working population in Germany, we find that the gig workers we study are substantially more risk-tolerant, on average, than the general population and even more risk tolerant than self-employed

[3] See, for instance, "Fashion forward - One of Europe's most interesting technology companies sells shoes and threads" in: The Economist, September 1st, 2016.

individuals. In turn, the loss in expected earnings per customer may have counterbalanced the insurance effects.

Our findings contribute to a nascent literature on the gig economy in applied microeconomics. Hall and Krueger (2018) present a detailed characterization of Uber's U.S. driver-partners documenting, for instance, their greater similarity to the general workforce than to taxi drivers. The authors identify the flexibility to set hours as a key determinant of driver-partners' decision to work with Uber.

Angrist et al. (2017) use an RCT at Uber to study the value to ride-hailing drivers of working with Uber as compared to a scenario similar to working as an independent taxi driver. Their experiment offers Uber drivers the opportunity to reduce the farebox share they have to pay Uber by purchasing a virtual taxi medallion lease. They use drivers' self-selection into the treatment to quantify how valuable it is for gig workers to be able to drive without a lease. Based on this they argue that Uber's existence creates a welfare gain for would-be entrepreneurs who would not purchase a taxi medallion. Our paper is similar to Angrist et al. (2017) in that both studies experimentally vary the incentive intensity on a gig-economy platform and study the response of freelancers' labor supply. Our focus is very different, though: while they estimate the value of a specific gig-economy platform to its freelancers, we investigate costs and benefits of insuring its gig workers against income uncertainty.[4]

We further contribute to a small body of work showing that personality traits may substantially affect the impact of performance pay on individual performance. Callen et al. (2015) study the effect of a monitoring intervention among health inspectors in Pakistan and find that the intervention raised performance to a stronger extent for inspectors with more pronounced personality traits.[5] Donato et al. (2017) study heterogeneous treatment effects of performance incentives offered to obstetric providers with respect to personality traits and show that performance incentives raise health outcomes significantly only for less conscientious providers and those with high emotional stability. The results of this small literature all confirm our findings in that they also point to considerable heterogeneity in the effect of incentives depending on the incentivized individual's personality.

Finally, our results may have implications for policy makers. Among (mainly legal) scholars, lawmakers, regulators and in the public domain there is an ongoing debate about whether gig economy firms should be required to treat freelancers as regular employees (Means and Seiner (2015), Cherry and Aloisi (2017), Kuhn and Maleki (2017), Prassl (2018)).[6] This would have implications, for instance, for unemployment insurance, fair labor standards, and other labor market laws and regulations.[7] Businesses facing potential regulation

[4]Further recent papers on labor supply of gig workers include, for instance, Cook et al. (2018) who document a sizeable gender earnings gap among Uber drivers that is explained by gender differences in experience and preferences over the place and speed of driving. Stanton and Thomas (2019) study determinants of demand and supply on a large global online market place for gig workers exploring reasons for initially low adoption rates of outsourcing tasks to gig workers.

Our paper is also related to a larger number of recent papers which use online labor markets such as MTurk to recruit subjects for a task where experimenters vary payment schemes such as Burbano (2016), Burbano (2019), DellaVigna and Pope (2018b), DellaVigna and Pope (2018a), and List and Momeni (2019). However, in these papers workers are hired once for short term one-off tasks that typically take less than an hour to perform. We consider workers that repeatedly work on the task over a substantially longer period of two months in a natural work setting (rather than a setting that is created for research purposes). Additionally, earnings in these experiments are typically lower, whereas in our setting, median gig worker earnings amount to EUR 562.

[5]Callen et al. (2015) average scores of a big five personality test into one "Big 5 index" and find significant positive interaction effects of the intervention with that index. When considering the traits separately, they find weakly significant interaction effects for agreeableness, conscientiousness, and emotional stability.

[6]A visible example saw the European Court of Justice rule that Uber is a provider of transport services, not just an intermediary matching ride-hailers with self-employed drivers. See, for instance, https://www.bbc.com/news/business-42423627.

[7]Hagiu and Wright (2019) develop a formal economic model to analyze the implications of classifying workers as employees

might fear that such rules would drastically increase costs and reduce the performance of their workforces. If, for example, gig economy firms had to convert their freelancers into regular employees, they would probably have to comply with minimum wage provisions. While our intervention does not test the introduction of a minimum wage for freelancers, its fixed payment per order in effect introduces an earnings floor per unit of time the freelancer makes available. Consequently, our results may be useful for policy makers trying to assess the costs (through potential losses in performance) that stricter regulation would impose on gig economy firms.

2 Research setting

We collaborate with Zalon, an online provider of curated fashion shopping. Zalon is owned by Zalando, the leading online fashion retailer in Europe. Styling service, shipping and returns are free to customers, while being based on items from Zalando's online shop. The idea is that people, for various reasons, can delegate parts of the shopping process to a remote, professional stylist to receive specific outfit suggestions. Zalon is run as a platform: it relies on freelancers (the "stylists") to provide the styling service and matches clients and stylists according to their preferences.

To use Zalon's service, a client signs up on the company's website providing information about expectations, preferences, sizes and the willingness to pay. The prospective customer is then presented with a choice of two or three profiles of stylists matching her fashion preferences (see figure 6 in the appendix). Once a client has chosen a stylist, she sticks with her unless she actively picks a different stylist.

When a client wants to place an order, both parties are able to communicate to clarify, for instance, the occasion for which the outfit is intended. The stylist then puts together a set of items using a range of software tools provided by Zalon. She sends an outfit preview to the client, allowing for another feedback loop. Once the client has confirmed the order, Zalando packages and ships the outfit (see figure 7 in the appendix). The client pays for only the elements of the outfit that she decides to keep; everything else she returns, at no cost, with Zalon handling returned items.[8]

Stylists work remotely, typically from home. They decide how many slots to make available in an online calendar system for providing curated shopping services. Stylists are free to choose different slots every day and are encouraged to provide a weekly minimum number of slots (see figure 8 in the appendix). This minimum is not compulsory though. As a consequence, stylists can determine their labor supply rather flexibly. Stylists of course also influence sales on each order: the more carefully they match their combinations of items to the customer's taste and budget, the larger the sales resulting from the order will be.[9]

Prior to the intervention, Zalon's stylists' compensation was purely commission-based pay: stylists received a share of sales per order resulting from their curated shopping service. In the time frame considered, stylists in the control group received 15% of the value of the items kept by the client.[10] If the client did not keep anything the stylist had chosen, stylist earnings were zero under this scheme.

rather than independent contractors.

[8] Apart from software tools for creating outfits, Zalon also provides stylists with sales information on a monthly basis, allowing her to evaluate the performance of (combinations of) items she has chosen for clients.

[9] Zalon recruits its freelance stylists through a multi-step application procedure on a rolling basis. In addition to providing information on their qualifications and experience, applicants need to put together a trial outfit from Zalando's online shop (as they would do for prospective clients). Stylists are selected based on the quality of their application (evaluated by senior stylists). Once an applicant has been chosen, an offer is made that usually results in a contract.

[10] The commission rate was reduced to penalize the gig worker for specific undesired types of behavior like, for instance, frequent failure to respond to client messages.

3 Theoretical framework

Our hypotheses are based on a formal model aimed at capturing central features of both our field experimental setting and the gig economy more generally: agents' ability to choose both how much to work (number of assignments) and how much effort to invest into these assignments.[11] Our framework builds on a Holmström and Milgrom (1991)-type multi-tasking model. Consider an agent who works for a principal, providing a service to customers. The agent chooses the number of client orders to fulfill $n \in [0; \bar{n}]$ and the average service quality $q \in [0; \bar{q}]$. The agent has (potentially) imperfectly known ability $a \sim N\left(m, \sigma_a^2\right)$ with $m > 0$. In order to study comparative statics with respect to behavioral determinants of the agent's effort reaction consider the specific functional form of the cost function[12]

$$c\left(q, n\right) = n\left(\frac{\kappa}{2}q^2 - \frac{\eta}{2}\left(\tau - \left(q - q^*\right)^2\right)\right) + \frac{\nu}{2}n^2 \tag{1}$$

with $\eta, \kappa, \eta \in [0, \infty[$ and $\tau \in \left[0, q^{*2}\right]$ where q^* is the first-best level of quality. If $\eta = \tau = 0$, the agent is purely extrinsically motivated and $c\left(q, n\right) = n \cdot \frac{\kappa}{2}q^2 + \frac{\nu}{2}n^2$. In this case, the marginal cost of fulfilling another order as well the marginal cost of providing more quality per order are strictly increasing. Moreover, the cost of providing a quality level q on each order is increasing in the number of orders. If, however, $\eta > 0$, the agent is (at least to some extent) intrinsically motivated for the task and therefore has a preference for providing a quality level that is close to the normatively optimal (first best) level q^*.[13] The parameter τ shifts the intrinsic utility of serving a customer.[14]

When the agent fulfills n orders, she generates a level of sales

$$S = \sum_{i=1}^{n}\left(a + q + \varepsilon_i\right)$$

where $\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$. The agent has an outside option that yields a reservation certainty equivalent $w_A > 0$. We allow for the possibility that the agent is risk averse with constant absolute risk aversion, where her Arrow-Pratt measure of absolute risk aversion is r.

Both total sales S and the number of orders worked on n are verifiable and we consider linear contracts that pay a wage

$$w = \alpha + \beta \cdot n + \gamma \cdot S,$$

where $\beta \geq 0$ is an order bonus, i.e., an order-based piece rate that does not depend on quality, and $\gamma \in [0, 1]$ is

[11] We included our hypotheses in the pre-registration of the field experiment on the AEA's social science registry (AEARCTR-0001555). A few weeks after the start of the trial period but before obtaining access to any data, we uploaded the analysis of the formal model. Here we present the proofs of the four hypotheses we preregistered. The uploaded longer version of the theoretical analysis can be requested from https://www.socialscienceregistry.org/trials/1555 under "Supporting Documents and Materials".

[12] See our preregistered formal analysis for a more general version of the model.

[13] In the setting we study the agent picks a set of items that is sent to customers via mail. The customers can then decide which items to keep and enjoy free returns of the items they do not want. As the firm incurs costs without earning anything on all returned items its "ideal" agent only selects items the customer wants to keep. In this respect, the firm's objective function is closely aligned with the customer's interests. Formally, we thus define $(n^*, q^*) = \arg\max_{n,q} n\left(m + q\right) - \left(n\left(\frac{\kappa}{2}q^2 - \frac{\eta}{2}\left(\tau - \left(q - q^*\right)^2\right)\right) + \frac{\nu}{2}n^2\right)$ which yields quality level $q^* = \frac{1}{\kappa}$.

[14] Note that τ allows for different drivers of intrinsic motivation. If $\tau = 0$ the intrinsic motivation comes from the feeling of duty when having accepted to serve a customer. If, however, $\tau > 0$, it captures task enjoyment. Accepting an order gives the agent an opportunity to enjoy working. As will become clear below τ does not affect the agent's quality choice, but of course raises incentives to provide quantity.

a commission rate.

We first characterize an agent's reaction to a contract with a commission rate $\gamma \in [0,1]$ and an order bonus $\beta \geq 0$ obtaining that[15] the agent chooses a quality level

$$q = \frac{\gamma + \eta q^*}{\kappa + \eta}$$

and quantity

$$n = \frac{1}{\nu + r\gamma^2\sigma_a^2}\left(\beta + \gamma m + \frac{(\gamma + \eta q^*)^2}{2(\kappa + \eta)} - \eta\frac{q^{*2} - \tau}{2} - \frac{1}{2}r\gamma^2\sigma_\varepsilon^2\right).$$

Hence, quality q is unaffected by the order bonus β and increasing in the commission rate γ. But if the agent is intrinsically motivated quality is less responsive to γ (i.e. $\frac{\partial^2 q}{\partial\gamma\partial\kappa} < 0$), i.e. for intrinsically motivated agents a reduction in γ reduces q to a weaker extent. For $\gamma \to \infty$ quality q becomes inelastic in the commission rate γ. The chosen quantity n is increasing in the commission rate γ and the order bonus β.

We now develop predictions on the effect of the treatment intervention implemented in the field experiment. To this end, consider a shift from a pure commission rate $\gamma_0 \in]0,1]$ to a lower commission rate $\gamma_1 < \gamma_0$ combined with an order bonus $\beta > 0$; the relative size of $\gamma_0 - \gamma_1$ and β is calibrated on a population of agents in such a way that the average agent's pay per order remains constant if agents do not adjust quality. We now analyze the (heterogeneous) effects of such an intervention on expected quantity and quality assuming that the "personality traits" r_i and η_i are uncorrelated.

Proposition 1 *Consider an intervention that shifts compensation from a pure commission rate $\gamma_0 \in]0,1]$ to a lower commission rate $\gamma_1 < \gamma_0$ combined with an order bonus $\beta > 0$ that keeps the payment per order constant (at prior quality levels). This implies:*
(i) The intervention increases expected quantity ($E[\Delta n_i] > 0$).
(ii) The quantity increase will be the larger, the more risk averse the stylist is ($\frac{\partial E[\Delta n_i] > 0}{\partial r_i} > 0$).
(iii) The intervention will reduce expected quality $E[\Delta q_i] < 0$.
(iv) The quality loss will be smaller, the more intrinsically motivated an agent is (i.e., $\frac{\partial E[\Delta q_i|\eta_i]}{\partial \eta_i} > 0$).

Proof: See appendix.

The key intuition is the following. The intervention should raise *quantity* (i.e. the agent should offer more assignments) as the introduction of the order bonus which does not depend on the (uncertain) sales generated through the assignment provides a better insurance to the agents in their earnings from each assignment.[16] This effect is larger if the agent is more risk averse as risk averse agents benefit more from the better insurance and in turn should increase their quantity to a stronger extent.

[15] See the appendix for details.

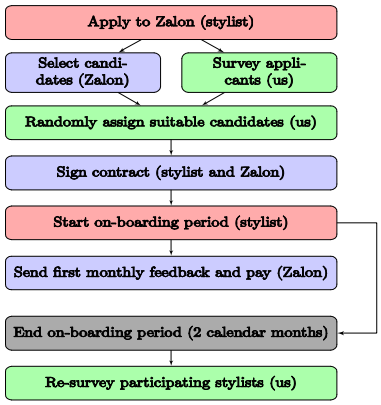
[16] The model captures the uncertainty associated with sales per order, but it does not capture a potential uncertainty associated with the number of customers actually assigned. In the field setting we study, stylists do not choose the number of client orders directly, but they commit to a number of time slots in which they are willing to serve customers. Hence, instead of directly choosing n, stylists choose an upper boundary for n. But note that when an offered time slot is not filled, stylists learn this in advance and also do not bear the costs for serving the customer. Only if the time slot is filled, the stylist actually has to work in that slot and then the insurance effect sets in. Hence, the key earnings uncertainty is driven by customers who send back items after the stylist has exerted effort on the recommended outfit.

However, the intervention may lower *quality* (i.e. sales per assignment) as the lower commission rate reduces incentives to generate sales from given assignments. This effect should be strongest for agents that show no intrinsic motivation for the task. For more intrinsically motivated agents this quality loss should be weaker. Intuitively, a motivated agent works more towards a normatively optimal quality for intrinsic reasons once having accepted an assignment.

4 Experimental design

The intervention randomly varied new stylists' pay. Stylists were unaware they were participating in an experiment, making experimenter demand effects unlikely. We restricted the experiment to new stylists as these had no previous experience with Zalon or its compensation policies.[17] This also allowed us to avoid spillover effects as each new stylist was individually informed about their respective compensation and as stylists work from home.[18] Figure 1 illustrates how our experiment integrates with Zalon's hiring process. The remainder of this section provides details on the intervention, data collection and group assignment.[19]

Figure 1: Experimental design and hiring process



4.1 Intervention

In the purely commission-based pay scheme (control group), earnings were 15% of the sales realized as a consequence of the gig worker's curated shopping service for a client. In the combined pay scheme (treatment group), the commission was 7.5% of sales per order and a fixed payment of €6.50 per confirmed order was added as an insurance component for the stylist. The fixed payment is due for each order confirmed by the client based on the outfit preview provided by the stylist, irrespective of sales or the number of returned items (that is, even when the customer sends all items back). The treatment was calibrated using pre-experimental

[17] Were we to include gig workers who had been working for the platform for a while and altered their incentives, the effect may be strongly influenced by existing reference points and responses to the perceived generosity or fairness of the change in pay.

[18] The claim that our setting was particularly well-suited to avoiding spillovers is supported by the fact that not a single person complained to Zalon about receiving different pay than a peer. Moreover, only one respondent mentioned in our post-treatment survey that she had heard of someone else with different pay. That comment, however, referred to sales commission rates that had been higher in the past and not to the different treatment.

[19] As profit margins of items bought are highly confidential Zalon was unable to provide information on profits, our third primary outcome variable.

sales data such that, assuming no quality adjustment, expected earnings per order would be similar under the two compensation schemes. The individual treatment period started after a newly hired stylist had been informed about the compensation scheme once they were free to start working. For a given new stylist, the treatment period ended after her first two calendar months with the platform (the so-called onboarding period). Treated stylists knew ex ante that the use of the fixed payment per order is restricted to two months and that they would be paid entirely by commission (15%) thereafter. Our intervention included all 202 stylists starting work for Zalon in Germany between October 2016 and August 2017.[20]

4.2 Data collection

We collected baseline data[21] using an incentivized online survey among all new applicants to the platform. Most importantly, the survey elicited measures for applicants' risk aversion and intrinsic motivation for the job using the following items: "How do you see yourself: are you generally willing to take risks or do you avoid taking risks?" (11-point Likert scale from 0 "not at all willing to take risks" to 10 "very willing to take risks") and "To what extent do the following two statements apply to you personally? 1) Turning a wallflower into a handsome person is much more important to me than earning a lot of money through my work. 2) It makes me particularly happy to find the perfect style for someone else even if that takes a lot of convincing and patience." (11-point Likert scale from 0 "does not apply at all to me" to 10 "fully applies to me"). The survey furthermore elicited household characteristics, education level, information on other jobs and earnings as well as personality traits. From an incentivized post-intervention online survey we obtained measures of job and pay satisfaction and turnover intention. Pay and job satisfaction both we measured using a single item "How satisfied were you in the first two months with your work at Zalon?" or "... with your personal income from your work at Zalon?", respectively (11-point Likert scale from 0 "completely dissatisfied" to 10 "completely satisfied"). Turnover intention was elicited using the single item "How many times since you joined Zalon have you thought about quitting your work as a stylist for Zalon?" (5-point Likert scale from 0 "never" to 4 "daily").[22]

Zalon provided detailed administrative data on stylists' performance during and after the treatment period, in particular measures of stylists' choices of labor supply (desired number of slots) and sales performance on each order (net merchandise value). For reasons of confidentiality, all outcome variables are normalized as percentages of the mean value of the control group.

4.3 Group assignment

Zalon recruits stylists on a rolling basis. Our setting thus required a randomization strategy that can quickly allocate individuals who enter at irregular frequencies, while balancing treatment group size and providing reasonable power. To this end we sequentially randomized successful applicants in pairs[23] within

[20] There was one person who started work without us knowing, so that we were unable to assign this individual and hence, excluded this individual from our analyses. There were another six people whom we assigned as planned but Zalon's HR department erroneously changed that assignment by informing them about the payment scheme they had not been assigned to. We also excluded these individuals from our analyses.

[21] The survey was conducted prior to allocation into treatment or control group.

[22] Both surveys were implemented using soscisurvey.de. See the online appendix for the German text of the questionnaires and an English translation.

[23] In the medical literature on clinical trials, where sequential randomization is commonplace due to patients trickling in, the method is known as (permuted) block randomization. See, e.g., Moore and Moore (2013) and Zelen (1974).

pre-defined strata of predicted labor supply. Power simulations prior to the intervention indicated that with this randomization technique our expected sample size of 200 stylists would give us a minimum detectable effect size for labor supply of .35 standard deviations with 79% power. For details on randomization protocol and power calculations see section 8.1 in the appendix.

Spillover or contamination effects are very unlikely in our setting for three reasons. First, stylists are not told they are part of an experiment. Second, we/our universities were in contact with participants only through online surveys which avoided reference to an experiment. Third, we had several ways of detecting suspicion on the part of the participants: first, they had the opportunity to comment on the composition of pay during their first two months with Zalon in our follow-up survey; in addition, Zalon agreed to inform us about any inquiries or complaints, but they did not get any. Spillover effects from one group to the other as participants find out about other stylists being exposed to a different pay structure are difficult to rule out completely. However, they are very unlikely in our setting: all stylists are newly hired and therefore unlikely to know each other, they come from all over Germany and they all work remotely. The first event for which they visit Zalon's headquarters in Berlin only happens after their first two calendar months (the treatment period). Finally, if a number of stylists had been aware of their peers being paid differently, it is likely some would have complained to Zalon or in the open-text section of our follow-up survey. However, no stylist made any such comments to Zalon and only one person mentioned in our follow-up survey that they knew of someone else with different pay. That comment, however, referred to sales commission rates that had been higher in the past (i.e., the respondent had presumably been encouraged to apply by an existing stylist).

5 Results

5.1 Descriptive statistics

187 stylists participated in our survey and we have administrative data on 202 stylists. The summary statistics in table 1 show that Zalon's stylists are relatively young and predominantly female. They are highly educated: around 70 percent have a high-school diploma (Abitur) and 42 percent have a university degree, with approximately 20 percent still in education. For around 90 percent of stylists this is not their only job. Before starting work there, stylists expect to work close to 16 hours a week for Zalon. Most stylists live with a partner and only a few have children.

Table 1 assesses the degree to which randomization achieved covariate balance between the 92 treatment stylists and the 95 control stylists. There is only one marginally significant difference, namely in the number of children. This deviation from perfect balance is consistent with a chance outcome. We also check covariate balance for personality traits. The treatment and control groups differ significantly (at the 10% level) only in their positive reciprocity, as table A.1 in the appendix shows. It is again consistent with random variation that out of eight t-tests, one delivers a non-zero difference significant at the 10% level. To conclude, we view the checks on covariate balance as evidence that randomization has been successful.

Table 1: Covariate balance: demographic characteristics

	All Mean	Control Mean	Treated Mean	Difference	(p-value)
Age (in years)	29.92	30.13	29.71	0.42	(0.648)
Male	0.09	0.09	0.09	-0.00	(0.965)
Abitur	0.70	0.68	0.72	-0.03	(0.622)
University	0.42	0.46	0.37	0.09	(0.196)
Hourly wage other jobs (in Euros)	29.35	27.77	30.97	-3.20	(0.533)
Monthly net income other sources (in Euros)	1,076.65	1,074.84	1,078.51	-3.67	(0.971)
Holds other job	0.91	0.94	0.88	0.06	(0.184)
Pursues education	0.20	0.18	0.23	-0.05	(0.406)
Expected weekly hours for Zalon	15.65	15.57	15.73	-0.16	(0.872)
Household size	2.22	2.19	2.26	-0.07	(0.638)
Number of children	0.23	0.31	0.15	0.15*	(0.062)
Observations	187	95	92	187	

5.2 Labor supply

We start with the experiment's average effect on stylists' labor supply. We measure labor supply by the number of desired slots during the treatment period - the slots stylists make available in their online calendars, normalized as percentages of the mean value of the control group.

Column (1) of table 2 reports results from a regression that includes only dummies for randomization strata and hire month as well as the number of days the individual was treated.[24] We find no statistically significant effect of the intervention on labor supply.[25] We next look at differences in treatment effect by risk aversion. We use our self-reported risk attitude item from the baseline survey and classify stylists as risk-averse (coded as 1) who are less willing to take risks than the median stylist, and 0 otherwise. Using this median split dummy we find no significant effect heterogeneity by risk aversion (see column (2) of table 2).[26] Our model predicts that more risk averse stylists would increase their labor supply more (hypothesis 2), but this is clearly not the case. If anything, the negative point estimate of the interaction term points into the opposite direction. Hence, the absence of the hypothesized effect is unlikely to be due to a lack of statistical power.

[24]The treatment lasted for two calendar months for all stylists, resulting in different treatment duration depending on exact starting date. Treatment duration varied from 31 to 62 days.

[25]This result is robust to including additional control variables as well as non-parametric estimation methods such as a Mann-Whitney U-Test (p-value = 0.623). Moreover, there is one obvious outlier. Removing it increases precision, but the estimated effect remains statistically insignificant.

[26]Using an experimentally elicited, incentivized measure of risk attitude instead leaves this result unchanged.

Table 2: Treatment effect on labor supply

	(1)	(2)
Treated	1.39	6.42
	(16.682)	(28.492)
Treated × risk averse		-25.65
		(36.391)
Risk averse		-1.71
		(17.243)
Adjusted R-squared	0.056	0.038
Number of observations	202	187

Note: Heteroskedasticity-robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Outcome variable: normalized stylist-level total number of desired slots. Controls: randomisation stratum, hire month and treatment duration. Risk averse is a median split dummy from the baseline survey measure (1 item). F-test treated + treated*risk-averse: p = 0.334.

5.3 Sales performance

Our formal model predicts a negative sales performance effect of the insurance treatment (hypothesis 3), which we expect to be weaker for stylists who are to a stronger extent intrinsically motivated for the task (hypothesis 4). To measure intrinsic motivation, we use a median split dummy based on the average of our two items on intrinsic motivation for the task from the baseline survey. We measure sales performance by net merchandise value (NMV, also called sales) per order in Euros, i.e., the value (after tax) of the items in a box that a customer keeps and pays for. Again, we normalize NMV as percentages of the mean value of the control group. Column (1) of table 3 reports an estimate for the average treatment effect on NMV per order from an order-level regression that includes only dummies for randomization strata and the timing of the order.[27] There is no statistically significant average treatment effect on stylists' sales performance.[28] However, the lack of an average treatment effect masks considerable effect heterogeneity by intrinsic motivation: In line with our hypothesis, the intervention reduces the sales performance of stylists with low intrinsic motivation, as column (2) of table 3 shows.[29] Moreover, the coefficient on the interaction term between the treatment and the intrinsic motivation dummy is positive and significant. As figure 2 illustrates, the treatment even tends to increase sales performance. The sales performance in this group of more intrinsically motivated stylists is weakly significantly larger than that in the control group (i.e. for the sum of treatment and interaction term, p = 0.099).[30]

[27] Timing controls are dummies for calendar weeks as well as a 3rd-degree polynomial in the days since starting to work for Zalon.

[28] Regression analyses that control for stylist characteristics confirm this result.

[29] Using the intrinsic motivation index as a continuous variable rather than a median split produces similar results, see table A.3 in the appendix.

[30] The sum of treatment and interaction term remains significantly different from zero at the 10% or 5% level in specifications that control for stylist and client characteristics, see table A.2 in the appendix.

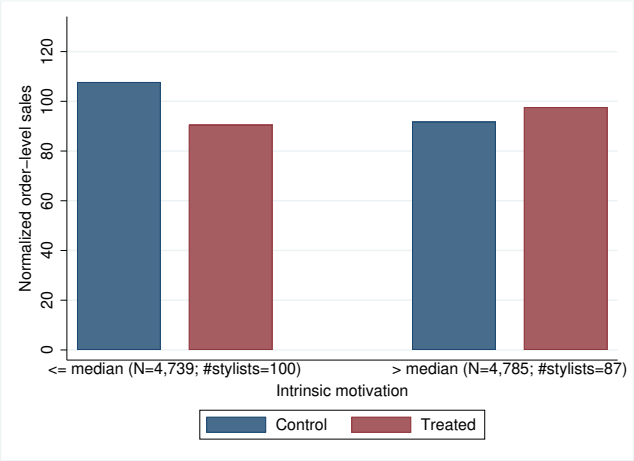


Figure 2: Effects on sales performance by intrinsic motivation

The heterogeneous effect estimates are not only statistically, but also economically significant. For stylists with intrinsic motivation weakly smaller than the median, our regression estimates imply that the treatment reduced (normalized) sales per order by 17.14%. The point estimate for heterogeneity by intrinsic motivation is also substantial, implying an increase in sales per order of 9.6% among the more intrinsically motivated stylists.

Table 3: Treatment effect on sales performance

	(1)	(2)
Treated	-4.43	-17.14***
	(3.946)	(5.686)
Treated × intrinsically motivated		26.70***
		(8.121)
Intrinsically motivated		-20.57***
		(5.747)
Adjusted R-squared	0.004	0.007
Number of observations	10,090	9,524
Number of stylists	202	187

Note: Standard errors clustered at stylist level in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Outcome variable: normalized order-level sales. Controls: randomisation stratum, calendar week and potential experience (3rd-degree polynomial). Intrinsic motivation (2-item index) is a median split dummy from the baseline survey measure. F-test treated + treated*intrinsically motivated: p = 0.099.

Finally, note that the more intrinsically motivated stylists' sales performance is lower under the high-powered incentive scheme with the full commission rate (i.e. in the control group). A potential reason for this is that intrinsically motivated stylists might be less keen to maximize sales and their earnings. In line with this conjecture we observe that among treated stylists, the sales performance of the more intrinsically motivated is not lower than that of the less intrinsically motivated stylists (the point estimate is positive but not significantly different from zero (p = 0.296)). We discuss this in more detail below.

We do not present detailed results on the effect heterogeneity by intrinsic motivation on labor supply as we did not pre-register a hypothesis on this. Still it is interesting to observe that the labor supply effects mirror the effects on sales performance: Labor supply is significantly lower under the treatment for the less motivated workers, but this is not the case for the more motivated ones. Labor supply even tends to be higher for the more motivated workers.

6 Discussion

6.1 Why is there no positive effect on labor supply?

Risk attitude In order to understand the absence of a positive effect on labor supply it is instructive to explore the risk tolerance of the individuals in our sample. Arguably, anticipating uncertain income streams less risk averse people may self-select into gig worker jobs such as the job we consider. If this is the case, the insurance effect may be rather weak even within the group of workers that are relatively more risk averse relative to the other workers on the platform: If average payments per gig are not higher in the treatment group, workers would then naturally not have an incentive to increase labor supply substantially.[31]

We can explore this conjecture by comparing the risk attitude among the workers we consider with the risk attitude in the general population of workers. The reason is that we have used identical items to assess risk attitude as are used in general population surveys. Table 4 compares responses of our stylist survey to data of the German Socio-Economic Panel (SOEP)[32], a representative, longitudinal data set of private households in Germany (Goebel et al. (2018)). In detail, we compare our gig workers workers with self-employed individuals and freelancers (except self-employed farmers) in the SOEP from the waves 2016 and 2017. The comparison of means shows that our stylists are indeed significantly more willing to take risks than self-employed individuals and freelancers in the SOEP population survey.[33] The stylists' low risk aversion may thus help to explain why our experiment did not significantly increase labor supply.

Table 4: Characteristics of stylists and private-sector employees

	Zalon stylists	SOEP freelancers		
	Mean	Mean	Difference	(p-value)
Risk attitude (1-11)	7.57	6.71	-0.86***	(0.000)
Observations	187	2,728		

* p<0.10, ** p<0.05, *** p<0.01

Lower pay in treatment group But in order to shed more light on his conjecture it is also instructive to explore the effect of the treatment on average earnings.[34] As our formal analysis shows, labor supply should

[31] But it seems not clear from the outset that gig workers are more risk tolerant than the general population as some may be driven to choose freelance work, for instance, because they want to finance their education or are forced by private circumstances to work from home.

[32] Socio-Economic Panel (SOEP), data for years 1984-2017, version 34, SOEP, 2019, doi:10.5684/soep.v34

[33] They also see themselves as significantly more conscientious.

[34] Earnings were a pre-registered secondary outcome.

increase for risk averse individuals when expected earnings per gig are identical in the treatment and control group or at least not much lower. Recall that the payment scheme in the treatment group was calibrated in a way that the average order would pay treated and control stylists similar amounts while varying the structure of pay: pure commission for control stylists, fixed payment plus lower commission for treated stylists.

But in the treatment phase, average earnings in the treatment group were lower than average earnings with the pure commission earned by control stylists.[35] This is illustrated by Figure 3. Because treated stylists' overall earnings disadvantage includes an (endogenous) performance response, the middle bar singles out the (exogenous) wage component.[36] Figure 3 makes it clear that the bigger part of the average treated stylist's earnings disadvantage was due to lower wages conditional on performance.

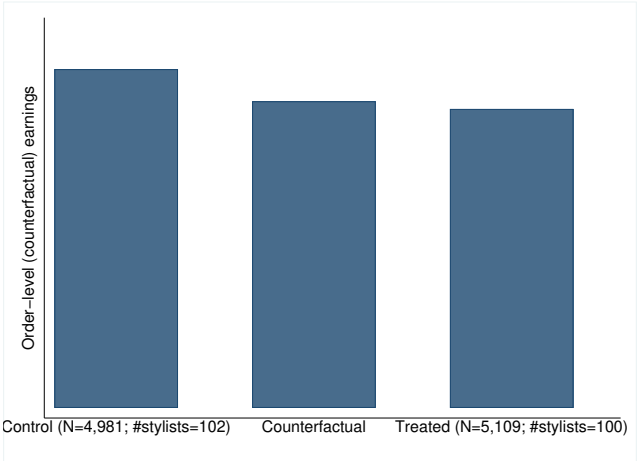


Figure 3: The wage component of the treatment effect on earnings

In fact, our formal model implies that stylists may have still benefited from the treatment (even though average wages are reduced) if the loss in average wages is offset by a utility gain from better insurance. But for one, as we have seen in the above, the gig workers we consider have rather high degrees of risk tolerance. Moreover, studies in comparable contexts have found positive wage elasticities of labor supply - see, e.g., Fehr and Goette (2007), Chen and Horton (2016), or Angrist et al. (2017). These results support the case for countervailing effects resulting in no significant overall effect of our intervention: lower wages depress the labor supply of freelancers, while the insurance effect of the treatment pulls in the opposite direction. Even though the low degree of risk aversion among the stylists apparently did not lead to a sufficiently strong utility gain to lead them to increase their labor supply, the insurance effect of the intervention may have avoided a substantial reduction in labor supply even though earnings per gig were reduced.

A second piece of evidence for the countervailing effects interpretation is that while the intervention reduced treated stylists' pay satisfaction, it had no significant effect on stylists' job satisfaction. These results from the post-treatment survey are presented in Table 5.

[35] The calibration was based on stylists' performance data from a summer month when seasonality caused sales to be below average, resulting in lower pay for the treatment group.

[36] The first and third bars give control and treatment stylists' earnings per order respectively. The middle bar shows the counterfactual earnings of hypothetical stylists who were paid less (than the control group) but did not change their behavior. That is, the middle bar shows what control stylists, holding performance constant, would have earned had they been paid according to the treatment's combined pay scheme.

Table 5: Treatment effect on satisfaction measures

	(1) Pay satisfaction	(2) Job satisfaction
Treated	-0.944**	-0.179
	(0.373)	(0.342)
Total earnings treatment period (100 EUR)		
Outcome mean	0.494	0.429
Adjusted R-squared	0.022	-0.002
Number of observations	168	168

Note: Heteroskedasticity-robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Outcome variable: stylists' responses to post-treatment survey items. Controls not shown: randomisation stratum, hire month and treatment duration.

6.2 How robust is the effect heterogeneity by intrinsic motivation?

Validating our measure of intrinsic motivation First, we use the post-intervention survey to explore whether our intrinsic motivation index based on stylists' responses before starting work for Zalon indeed measures intrinsic motivation. In line with this idea, stylists that claimed to be more intrinsically motivated before they started to work indeed score significantly higher on a seven-item task enjoyment index measured after two month of working[37] and report markedly lower turnover intention (see table 6). Their job satisfaction is also higher, though the difference in means is not statistically significant. In summary, stylists respond in ways consistent with differences in intrinsic motivation in a separate survey several months later, without being primed on the original questions or their answers from the baseline survey.

Table 6: Post-treatment survey responses by intrinsic motivation

	IM<=Median	IM>Median	Difference	
	Mean	Mean	Mean	(p-value)
Task enjoyment index (1-11)	8.20	9.14	0.94***	(0.000)
Job satisfaction (0-10)	7.53	7.92	0.39	(0.263)
Turnover intention (0-4)	0.74	0.38	-0.36**	(0.010)
N	87	77	164	

Intrinsically motivated stylist characteristics There is a justified concern that our measure for intrinsic motivation may simultaneously be capturing a range of other things. To explore this concern we start with a comparison of more and less intrinsically motivated stylists. Table 7 splits up the sample using the intrinsic motivation dummy from section 5.3 and tests for differences in characteristics measured in the baseline survey. The key result from this table is that intrinsically motivated stylists are different in two dimensions: their

[37] We apply the 7-item interest/enjoyment subscale from the intrinsic motivation inventory (IMI), a self-report measure of intrinsic motivation (Ryan (1982)). Like in, for instance, McAuley et al. (1989), the label "activity" has been reworded to reflect the nature of the task of our fashion gig workers.

personality and their education/earnings potential. We use these facts as starting points for our analyses in this section.

Table 7: Demographics and personality by intrinsic motivation (IM)

	IM<=Median	IM>Median	Difference	
	Mean	Mean	Mean	(p-value)
Age	30.04	29.78	0.26	(0.781)
Male (0/1)	0.07	0.13	-0.06	(0.202)
University entrance diploma (Abitur, 0/1)	0.75	0.64	0.11	(0.117)
University degree (0/1)	0.48	0.34	0.14*	(0.061)
Hourly wage other jobs	35.44	22.35	13.09***	(0.007)
Monthly net income	1,062.32	1,093.19	-30.86	(0.761)
Other jobs (0/1)	0.92	0.90	0.02	(0.583)
In education (0/1)	0.20	0.21	-0.01	(0.908)
Expected work hours (during onboarding)	15.03	16.36	-1.33	(0.192)
Household size	2.22	2.23	-0.01	(0.949)
Number of children	0.25	0.21	0.04	(0.602)
Risk attitude (survey item)	7.39	7.78	-0.39	(0.145)
Risk attitude (experiment)	12.10	10.19	1.91**	(0.027)
Intrinsic motivation (index)	7.34	9.89	-2.55***	(0.000)
Conscientiousness (index)	9.52	10.13	-0.61***	(0.000)
Time preferences (index)	7.41	8.18	-0.77***	(0.002)
Reciprocity (index)	4.43	4.82	-0.39***	(0.001)
Altruism item 1	111.71	180.11	-68.40***	(0.000)
Altruism item 2	8.56	9.70	-1.14***	(0.000)
Impulsiveness	6.53	6.60	-0.07	(0.808)
N	100	87	187	

Potential confounds: personality traits Unlike the treatment in our experiment, intrinsic motivation was not randomly assigned. As a consequence, it is possible that the effect heterogeneity by intrinsic motivation we find in fact captures heterogeneity by some omitted characteristic that is both correlated with intrinsic motivation and a determinant of sales performance. This may bias the coefficient of intrinsic motivation.

While it is impossible to control for unobserved omitted variables, table 7 identifies a range of personality characteristics that differ systematically by intrinsic motivation and that may also influence sales. The analyses presented so far do not account for these differences in stylist personality. To test whether these differences confound the observed intrinsic motivation effect heterogeneity, we include median split dummies for each of the relevant personality traits (risk attitude, conscientiousness, patience, reciprocity, and altruism) and their interactions with the treatment dummy in the sales performance regression from above. Table 8 shows that the coefficient on effect heterogeneity by intrinsic motivation remains similar in size and significance throughout this exercise, even when controlling for all the traits and interactions simultaneously. That including such a range of relevant characteristics does not much alter the picture gives us confidence that the heterogeneity by intrinsic motivation we find is not just capturing the effect of an omitted personality trait we do not observe.

Table 8: Personality-dependent treatment effects on sales performance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treated	-17.14***	-16.32**	-22.46***	-16.21***	-24.21***	-14.99**	-22.22**
	(5.686)	(7.575)	(6.574)	(6.020)	(6.031)	(6.382)	(8.692)
Treated × intrins. motivated	26.70***	29.06***	19.06**	27.15***	24.64***	29.52***	25.20**
	(8.121)	(8.543)	(8.152)	(8.563)	(8.468)	(8.300)	(9.803)
Intrinsically motivated	-20.57***	-21.46***	-19.89***	-21.64***	-20.99***	-21.44***	-22.49***
	(5.747)	(6.146)	(5.752)	(6.258)	(6.450)	(5.676)	(6.927)
Treated × risk averse (exp)		-1.68					-3.24
		(7.749)					(8.067)
Risk averse (experimental)		-2.57					0.74
		(5.613)					(6.308)
Treated × conscientious			17.87**				16.25*
			(8.861)				(9.644)
Conscientious			-4.62				-7.53
			(6.437)				(7.587)
Treated × patient				-1.17			-1.32
				(8.436)			(8.847)
Patient				3.74			3.35
				(6.015)			(6.492)
Treated × pos. reciprocal					13.34		13.11
					(8.375)		(8.734)
Positively reciprocal					0.80		0.54
					(6.126)		(6.874)
Treated × altruistic						-8.12	-13.28
						(8.616)	(8.846)
Altruistic						2.12	1.72
						(5.635)	(6.502)
Adjusted R-squared	0.007	0.008	0.008	0.007	0.008	0.007	0.010
Number of observations	9,524	8,950	9,524	9,524	9,524	9,524	8,950
Number of stylists	187	179	187	187	187	187	179

Note: Standard errors clustered at stylist level in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Outcome variable: normalized order-level sales. Personality characteristics are median split dummies from the baseline survey (indices or single items). Controls: randomisation stratum, calendar week and potential experience (3rd-degree polynomial).

Hence, the heterogeneous treatment effect with respect to intrinsic motivation seems very robust. Moreover, it is interesting to note that in particular conscientiousness leads to an independent heterogeneous effect in the same direction. If we consider our formal model, conscientiousness can be viewed as a trait that similarly affects work performance: a more conscientious person exerts a specific service level driven by a feeling of obligation which would also lead to a lower elasticity of sales to a reduction in marginal monetary incentives.

Potential confound: stylist productivity Table 7 shows that intrinsically motivated stylists systematically differ from their peers in their education and in the hourly wages they earn in other jobs. This could mean that intrinsic motivation is simply a proxy for lower stylist productivity and that the effect heterogeneity we attribute to differences in intrinsic motivation is instead due to ability differences. Several tests suggest this is not the case. We first check whether the treatment effect heterogeneity by intrinsic motivation persists when we include controls for stylists' university education and their hourly wages in other jobs (each interacted with the treatment dummy). As Table 9 shows, the treatment heterogeneity by intrinsic motivation remains similar in size and significance.

Table 9: Treatment effects on sales performance and outside options

	(1)	(2)	(3)	(4)
Treated	-17.14***	-15.70**	-15.03**	-14.22
	(5.686)	(7.219)	(7.498)	(8.645)
Treated × intrins. motivated	26.70***	26.41***	26.66***	26.58***
	(8.121)	(8.033)	(8.425)	(8.296)
Intrinsically motivated	-20.57***	-19.77***	-21.52***	-20.83***
	(5.747)	(5.718)	(6.036)	(6.003)
Treated × hourly wage other jobs		-0.06		-0.04
		(0.158)		(0.156)
Hourly wage other jobs		0.17		0.16
		(0.122)		(0.121)
Treated × university degree			-7.12	-6.84
			(8.700)	(8.169)
University degree			-3.50	-3.67
			(6.240)	(5.774)
Adjusted R-squared	0.007	0.007	0.007	0.008
Number of observations	9,524	9,524	9,524	9,524
Number of stylists	187	187	187	187

Note: Standard errors clustered at stylist level in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Outcome variable: normalized order-level sales. Information on outside hourly wage and university degree are from the baseline survey. Controls: randomisation stratum, calendar week and potential experience (3rd-degree polynomial).

To obtain a further measure of stylist productivity, we estimate stylist individual fixed effects from sales regressions using data for the two months after the treatment was completed, with everyone subject to the same pay scheme.[38] Using these estimated individual fixed effects as a productivity measure we check whether the effect heterogeneity by intrinsic motivation survives when we control for stylist productivity. Column (1)

[38] Our approach loosely follows Abowd et al. (1999) and Card et al. (2013). We regress order-level sales on dummies for individual week, calendar week, customer age category, customer country (Austria, Germany, Netherlands) and customer gender as well as absorbed stylist dummies, whose estimated coefficients we use as estimated stylist individual fixed effects.

of table 10 replicates the regressions testing for treatment effect heterogeneity by intrinsic motivation for the smaller sample of stylists for whom we have productivity estimates. [39] The results are very similar to those obtained before. Column (2) of table 10 introduces a median split dummy for the individual fixed effects. This variable is highly predictive of performance in the treatment period. However, it hardly changes the estimated effect heterogeneity by intrinsic motivation both in terms of size and significance. In column (3), we interact the treatment with the high-productivity dummy instead. If the heterogeneity we find were confounded by productivity, we would expect a similar result as with intrinsic motivation. This is not the case: the estimated effect heterogeneity is insignificant. This remains true if both interaction effects are included simultaneously, as can be seen in column 4.

Table 10: Effect heterogeneity by intrinsic motivation and stylist productivity

	(1)	(2)	(3)	(4)
Treated	-19.44***	-19.83***	-4.25	-16.61**
	(5.705)	(5.375)	(6.250)	(6.753)
Treated × intrinsically motivated	27.49***	26.61***		26.40***
	(8.409)	(7.677)		(7.637)
Intrinsically motivated	-20.15***	-19.37***		-19.16***
	(6.018)	(5.250)		(5.235)
High productivity		18.95***	23.05***	21.89***
		(3.830)	(5.888)	(5.543)
Treated × high productivity			-7.05	-5.59
			(7.855)	(7.484)
Adjusted R-squared	0.007	0.012	0.009	0.012
Number of observations	9,155	9,155	9,155	9,155
Number of stylists	167	167	167	167

Note: Standard errors clustered at stylist level in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Outcome variable: normalized order-level sales. Controls: randomisation stratum, calendar week and potential experience (3rd-degree polynomial). Intrinsic motivation (2-item index) is a median split from the baseline survey measure. High productivity is a median split of estimated stylist fixed effects from post-treatment sales regressions.

Taken together, we interpret our analyses that explicitly include an ability measure as evidence that while there is a connection between intrinsic motivation and performance, the treatment effect heterogeneity by intrinsic motivation we find is not confounded by stylist productivity.

Does the treatment pay more among intrinsically motivated stylists? As intrinsically motivated stylists have lower sales performance when there is a pure commission rate (see table 3), some of the effect heterogeneity we find may be due to the fact that for the more intrinsically motivated stylists, the treatment reduces their pay to a weaker extent. As a first step we repeat the thought experiment from section 6.1 separately for more and less intrinsically motivated stylists. Figure 4 suggests that the treatment did reduce wages for both extrinsically (p=0.000) and intrinsically motivated stylists (p=0.031), though this reduction was less pronounced among the intrinsically motivated. Figure 4 also illustrates that intrinsically motivated stylists limited their earnings loss as the treatment had a positive effect on their performance: earnings did not fall as

[39] For 20 stylists we do not have enough data to estimate the individual fixed effects from post-treatment sales.

much as the lower commission rate implies. By contrast, extrinsically motivated stylists' performance went down, exacerbating the negative effect on earnings of the treatment's lower pay. On the whole, we interpret figure 4 as further evidence that the intrinsically motivated stylists value the insurance effect of the treatment more and potentially also get less frustrated by lowered pay. This view is also supported by the post-treatment survey: there is no heterogeneity in the treatment's effect on pay satisfaction (see table A.4 in the appendix).

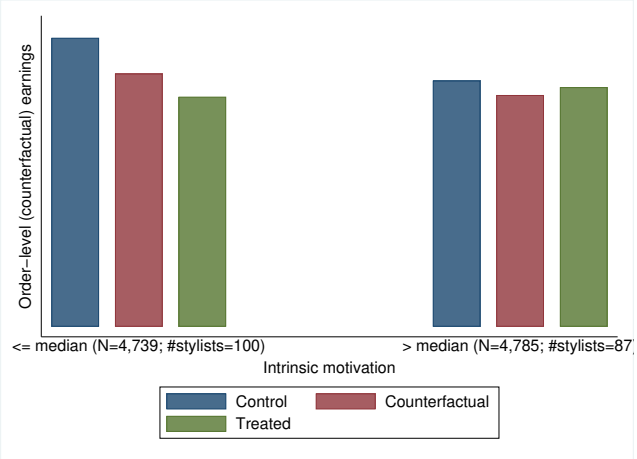


Figure 4: IM: CT earnings, counterfactual CT earnings, and TR earnings by IM

A note on intrinsically motivated stylists' lower sales performance

We have found that when insurance is provided through the treatment, the more intrinsically motivated agents achieve higher sales per gig as compared to the less motivated ones. But the opposite holds in the control group. That is, under pure commission-based pay, intrinsically motivated workers are less effective on the job. We have shown above that intrinsic motivation does not capture lower productivity. What then may explain this link between sales and intrinsic motivation under commission based pay? For one, intrinsically motivated stylists may have lower sales because, on average, they are less keen to maximize their earnings e.g. by selling more expensive items or they may be less "pushy" when attempting to get customers to re-order, e.g., by e-mailing them. While our data does not allow a direct test for this conjecture, we can perform an indirect test comparing the average value of items chosen by more and less intrinsically motivated stylists. When we regress the value of kept items[40] on the intrinsic motivation dummy, the estimated coefficient does point in the expected direction, though it is not statistically significant (p=0.193, Table A.5 in the appendix).

Alternative performance measure: Reputation and repeat orders

As a final piece of evidence to corroborate our finding we present regressions using another performance measure based on Zalon's order-level administrative data: the probability that a customer returns to order another outfit from the same stylist as a measure of a stylist's reputation.[41] For each stylist we calculate the

[40] We have data only on the value of items the client kept, not the ones they returned.

[41] This performance measure is known as repurchase rate in marketing and constitutes a commonly used metric in practice, particularly in retailing (Vogel et al. (2008); Anselmsson and Bondesson (2015)). An alternative measure, which is not based on decisions by past customers, would be the likelihood that a stylist is chosen from the list of three stylists proposed to any

proportion of customers who place more than one order during the first four calendar months.[42] We repeat our analyses of treatment effect heterogeneity by intrinsic motivation. Figure 5 confirms our findings from above: the treatment effect on less intrinsically motivated stylists is negative, while that on more intrinsically motivated ones is positive.[43]

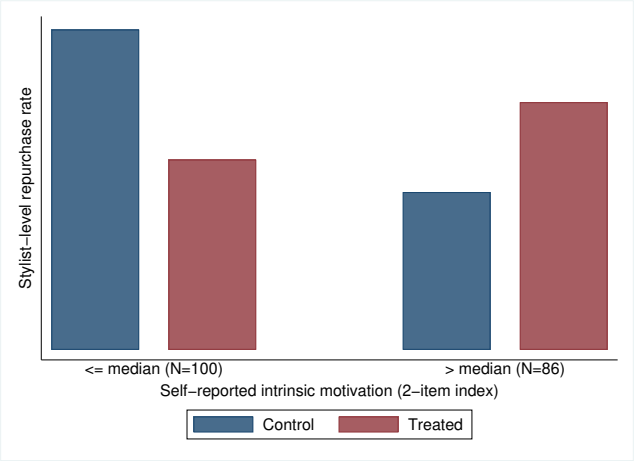


Figure 5: Effect on repeat orders by intrinsic motivation

Table 11 presents regression results, where (again for reasons of confidentiality) we normalize repurchase rates as percentages of the mean value of the control group. We find economically sizeable effects: for less intrinsically motivated stylists, the treatment leads to a reduction in repeat orders of 53%, whereas for the more intrinsically motivated stylists, we observe a 41% increase in the repurchase rate.

new customer. Unfortunately, we have no access to this selection as we have data only on matched stylists. But we believe that the decision of a repeat order by a former customer is the more robust measure as it is based on past quality rather than mere impressions from profiles of the stylists displayed to new customers.

[42] Because very few customers order more than once in two calendar months (the duration of the onboarding period), we use not only our treatment period, but add the two subsequent months.

[43] We also have some data on customer satisfaction. However, customers provide the likelihood they would recommend Zalon in less than 10% of cases in our data and the distribution of responses is extremely left skewed, suggesting that respondents are highly positively selected. Unsurprisingly, we find neither evidence for a treatment effect nor for effect heterogeneity using these data.

Table 11: Treatment effect on repeat orders

	(1)	(2)
Treated	-12.297	-52.764**
	(13.508)	(22.475)
Treated* × intrinsically motivated		94.019***
		(31.317)
Intrinsically motivated		-62.859***
		(23.378)
Adjusted R-squared	0.032	0.064
Number of observations	201	186

Note: Heteroskedasticity-robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Outcome variable: normalized stylist-level share of repeat customers (treatment + 2 months). Controls: randomisation stratum, hiring month and treatment duration. Intrinsic motivation (2-item index) is a median split from the baseline survey measure. F-test treated + treated*intrinsically motivated: p = 0.020.

7 Conclusion

This paper presents evidence from a randomized controlled trial on the effect of insuring gig workers against earnings risk. The treatment insured workers by reducing their sales-based commission while adding a fixed payment per order. The field experiment was implemented in collaboration with Zalon, an online provider of curated fashion shopping. 202 freelancers were part of the experiment for their first two calendar months with Zalon. If assigned to the control group, they were subject to the status-quo structure of pay, namely a pure commission (15% of sales per order). If assigned to the treatment group, they received a lower commission (7.5%) and a fixed payment of €6.50 per order.

We find that the insurance provided by the treatment did not increase labor supply. In contrast to the prediction based on standard economic reasoning, the treatment's lower-powered incentives also did not significantly lower performance. This is due to the fact that a significant loss in sales performance of the less intrinsically motivated stylists is offset by a significantly positive gain for the motivated ones. If we consider the 50% of stylists who did the job more for the money's sake than for the task's, the standard prediction holds: the treatment led to a substantial reduction in sales per customer and in the share of customers who ordered another outfit in a four-month observation window. Among the more intrinsically motivated half of stylists, on the other hand, the treatment had a positive effect on sales performance.

As intrinsic motivation was not randomly assigned, we cannot claim that this effect heterogeneity solely reflects a difference in how more intrinsically motivated stylists respond to the treatment. However, we show in a series of robustness checks that our finding is not driven by systematic differences in other personality traits nor by observed and unobserved ability. This rules out the most plausible alternative drivers of our result.

One caveat is that our study had enough power to detect large effects on labor supply (approximately a third of a standard deviation), but not small ones. Beyond that, the most convincing explanation for why we do not find a positive effect on labor supply is that pay in the treatment group not only provided insurance against income uncertainty, but was also lower on average. As the pool of stylists is substantially more risk tolerant than the general population, a positive insurance effect may not have been strong enough to increase

labor supply but it still has prevented a reduction.

This study has several implications. First and foremost, it shows that the effect of incentive schemes hinges on specific traits and preferences of the workforce. Providing better insurance through lower-powered incentives can (in line with standard theory) reduce incentives for more extrinsically motivated workers. But this loss can be offset by an increase in motivation for motivated workers.

Our findings on the effects of insuring freelancers against earnings risk are also relevant for other gig economy firms, many of which share characteristics of the setting we have studied: virtual platforms match workers with customers, the workforce is composed of freelancers rather than employees and freelancers' freedom to set their own schedules is emphasized, compensation is purely output-based, exposing the freelancers to substantial earnings risk. As our results emphasize the importance of the freelance workforce's motivation for the service offered, their implication for platforms in the on-demand economy is not uniform. When freelancers perform tasks they care about, it may be beneficial for gig economy firms to insure their workforces against earnings uncertainty. When the tasks attract gig workers primarily motivated by making money, our results suggest that the loss from reducing monetary incentive intensity outweighs any potential gains from insuring freelancers.

The results of our study may also inform the recent debate among policy-makers on the regulatory treatment of gig workers (see e.g. Means and Seiner (2015), Cherry and Aloisi (2017), Kuhn and Maleki (2017), Prassl (2018)). For instance, our findings indicate that the introduction of wage floors can hurt performance in areas where gig workers mostly work for the money and are less intrinsically motivated for the task. However, in areas where task motivation is large, the reduction of formal incentives induced by guaranteeing a minimum income per gig may not come with a loss in performance.[44] For more intrinsically motivated workers being provided with a better insurance against income uncertainty can even raise performance. Relatedly, our finding of positive effect heterogeneity by intrinsic motivation point to an adjustment strategy for employers relying mostly on freelancers, should regulation force them to introduce earnings floors: in this case stronger investments in the screening of gig workers with respect to their motivation for the respective task should become much more important.

[44] Our results thus complement the recent evidence on labor supply effects of the introduction of wage floors for gig workers by Horton (2018). He finds that wage floors led to a reduction in hiring of less productive workers. We do not study effects on the selection of workers (in our setting workers learned about the details of the compensation scheme after having been hired), but our setting allows to study effects of the change in the compensation scheme on performance of the hired workers.

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Abraham, K., J. Haltiwanger, K. Sandusky, and J. Spletzer (2017). Measuring the gig economy: Current knowledge and open issues. In C. Corrado, J. Miranda, J. Haskel, and D. Sichel (Eds.), *Measuring and Accounting for Innovation in the 21st Century*, NBER Book Series Studies in Income and Wealth. University of Chicago Press.

Angrist, J. D., S. Caldwell, and J. V. Hall (2017). Uber vs. taxi: A driver's eye view. Technical report, National Bureau of Economic Research.

Anselmsson, J. and N. Bondesson (2015). Brand value chain in practise; the relationship between mindset and market performance metrics: A study of the swedish market for fmcg. *Journal of Retailing and Consumer Services 25*, 58–70.

Burbano, V. (2019). Getting gig workers to do more by doing good: Field experimental evidence from online platform labor marketplaces. *Organization & Environment*.

Burbano, V. C. (2016). Social responsibility messages and worker wage requirements: Field experimental evidence from online labor marketplaces. *Organization Science 27*(4), 1010–1028.

Callen, M., S. Gulzar, A. Hasanain, Y. Khan, and A. Rezaee (2015). Personalities and public sector performance: Evidence from a health experiment in pakistan. Technical report, National Bureau of Economic Research.

Card, D., J. Heining, and P. Kline (2013). Workplace heterogeneity and the rise of west german wage inequality. *The Quarterly Journal of Economics 128*(3), 967–1015.

Chen, D. L. and J. J. Horton (2016). Are online labor markets spot markets for tasks? a field experiment on the behavioral response to wage cuts. *Information Systems Research 27*(2), 403–423.

Cherry, M. A. and A. Aloisi (2017). Dependent contractors in the gig economy: A comparative approach. *American University Law Review 66*(3), 635.

Cook, C., R. Diamond, J. Hall, J. A. List, and P. Oyer (2018). The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers. *Stanford Graduate School of Business Working Paper No. 3637*.

DellaVigna, S. and D. Pope (2018a). Predicting experimental results: Who knows what? *Journal of Political Economy 126*(6), 2410–2456.

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

Donato, K., G. Miller, M. Mohanan, Y. Truskinovsky, and M. Vera-Hernández (2017). Personality traits and performance contracts: Evidence from a field experiment among maternity care providers in india. *American Economic Review 107*(5), 506–10.

Fehr, E. and L. Goette (2007). Do workers work more if wages are high? evidence from a randomized field experiment. *American Economic Review 97*(1), 298–317.

Goebel, J., M. Grabka, S. Liebig, M. Kroh, D. Richter, C. Schröder, and J. Schupp (2018). The german socio-economic panel (soep). *Journal of Economics and Statistics 239*(2), 345–360.

Hagiu, A. and J. Wright (2019). The status of workers and platforms in the sharing economy. *Journal of Economics & Management Strategy 28*, 97–108.

Hall, J. V. and A. B. Krueger (2018). An analysis of the labor market for uber's driver-partners in the united states. *ILR Review 71*(3), 705–732.

Holmström, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership and job design. *Journal of Law, Economics and Organization 7*, 24–52.

Horton, J. J. (2018). Price floors and employer preferences: Evidence from a minimum wage experiment. *Working Paper*.

Kässi, O. and V. Lehdonvirta (2018). Online labour index: measuring the online gig economy for policy and research. *Technological forecasting and social change 137*, 241–248.

Katz, L. F. and A. B. Krueger (2019). The rise and nature of alternative work arrangements in the united states, 1995-2015. *ILR Review 72*(2), 382–416.

Kuhn, K. M. and A. Maleki (2017). Micro-entrepreneurs, dependent contractors, and instaserfs: Understanding online labor platform workforces. *Academy of Management Perspectives 31*(3), 183–200.

List, J. and F. Momeni (2019). Leveraging Upfront Payments to Curb Employee Misbehavior: Evidence from a Natural Field Experiment. Natural Field Experiments 00665, The Field Experiments Website.

McAuley, E., T. Duncan, and V. V. Tammen (1989). Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport 60*(1), 48–58.

Means, B. and J. A. Seiner (2015). Navigating the uber economy. *UCD Law Review 49*, 1511–1546.

Moore, R. T. and S. A. Moore (2013). Blocking for sequential political experiments. *Political Analysis 21*(4), 507–523.

Prassl, J. (2018). *Humans as a service: The promise and perils of work in the gig economy*. Oxford University Press.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology 43*(3), 450–461.

Stanton, C. T. and C. Thomas (2019). Missing trade in tasks: Employer outsourcing in the gig economy. Technical report, Centre for Economic Performance, LSE.

Vogel, V., H. Evanschitzky, and B. Ramaseshan (2008). Customer equity drivers and future sales. *Journal of marketing 72*(6), 98–108.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases 27*(7), 365–375.

8 Appendix

Figure 6: Stylist profiles

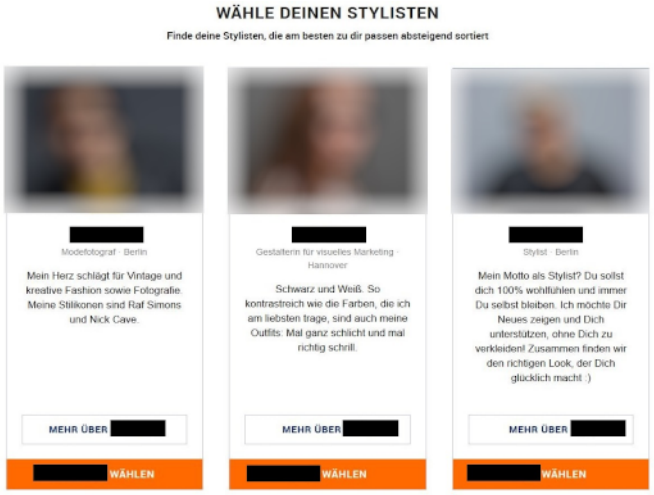


Figure 7: Sales process

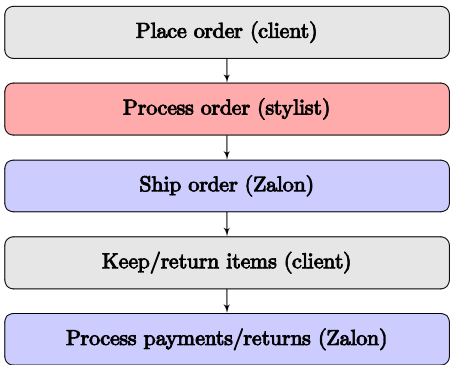
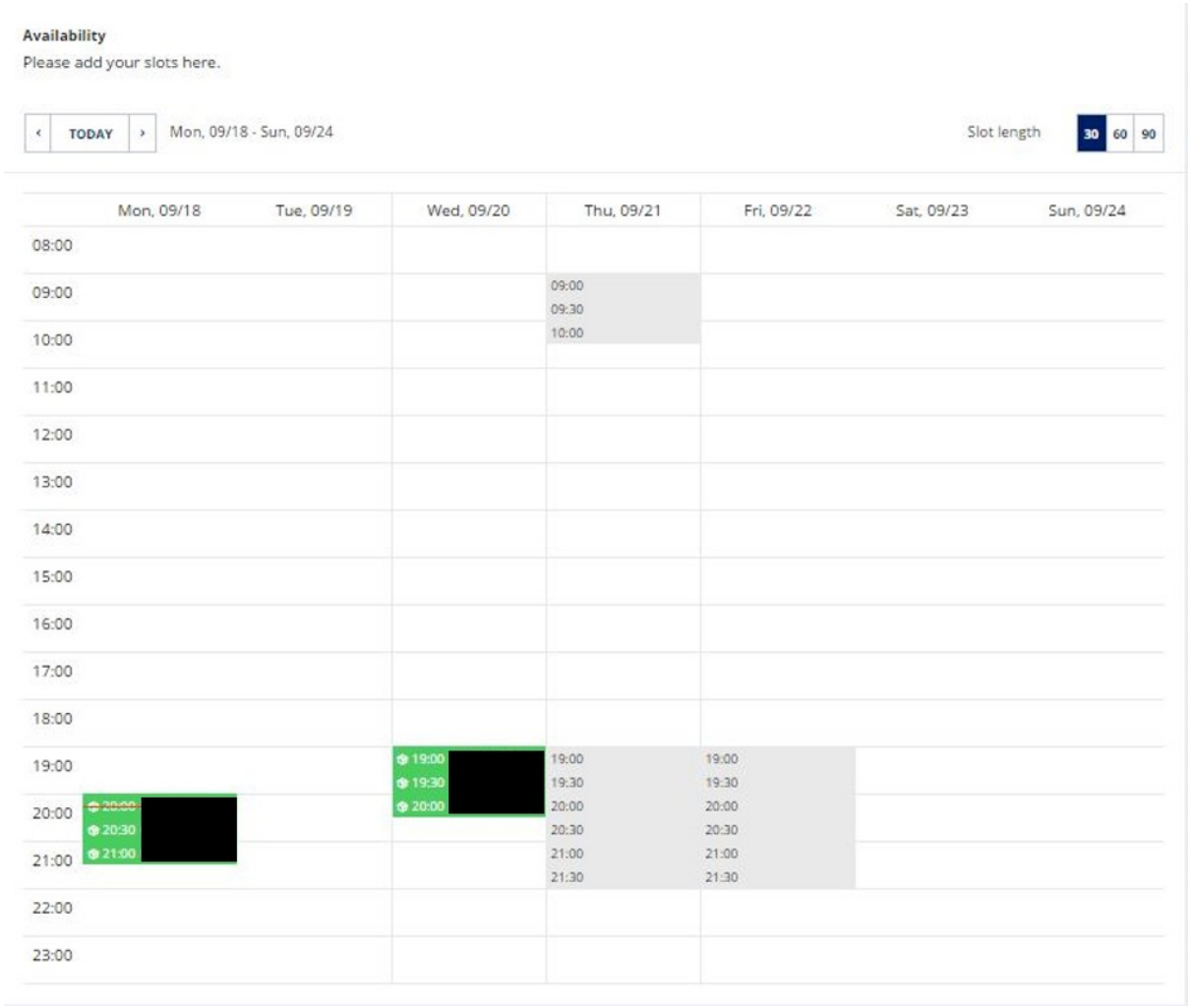


Figure 8: Stylist calendar



8.1 Randomization and power

To implement sequential randomization in pairs within pre-defined strata of predicted labor supply, we proceeded as follows: when Zalon informed us that a new stylist had been selected to receive an offer, we first made a prediction of the stylists's labor supply (number of slots).[45] We next checked into which one of the four predicted activity strata the prospective stylist's number of slots falls.[46] We then randomly allocated prospective stylists in pairs within their stratum: if the number of treated individuals in the stratum equals

[45] For this prediction, we use information from the baseline survey on intended activity levels, age and number of children, which we feed into predictive regressions. Before the start of the experiment, we had drawn a random sample of existing stylists to participate in (a slightly modified version of) our baseline survey for prospective stylists. This training sample was used to choose the regression specification that best predicted their actual labor supply. We then used these results to obtain the predicted labor supply for new stylists.

[46] We also used the training sample (N=86) to define predicted labor supply strata: stratum 1 (up to the 30th percentile), stratum 2 (p30 to p60), stratum 3 (p60 to p90) and stratum 4 (above p90). Stratum 5 includes individuals for whom we lack predicted labor supply as they did not participate in the baseline survey.

the number of control individuals, we tossed a fair coin. If the number of treated and control individuals was different, the stylist was the second in her pair and allocated so as to equalize group sizes. This ensures that within strata, group sizes remain approximately balanced at any given time. In addition, stratifying by predicted labor supply (albeit imperfectly due to the sequential entry into the experiment) was intended to increase power for the detection of treatment effects on stylists' choice of labor supply.

Our choice of randomization strategy had partly been informed by our power simulations that approximated various randomization strategies. These simulations gave the following result for the randomization strategy we have chosen: with 257 observations (the number of active stylists for whom past data was available), our simulated minimum detectable effect size (MDE) for (actual) labor supply with 77% power was equal to 0.3 standard deviations.[47] Due to planning constraints, however, Zalon was able to make a commitment for only five months in the field (yielding an expected 200 observations).[48]

8.2 Results

Table A.1: Covariate balance: personality

	All	Control	Treated	Difference	
	Mean	Mean	Mean	Mean	(p-value)
Self-reported risk attitude (1-11)	7.57	7.45	7.70	-0.24	(0.363)
Experimentally elicited risk attitude (1-21)	11.22	10.66	11.80	-1.14	(0.183)
Intrinsic motivation (2 items, 1-11)	8.53	-0.07	0.07	-0.14	(0.340)
Conscientiousness (3 items, 1-11)	9.81	9.80	9.82	-0.02	(0.923)
Patience (2 items, 1-11)	7.77	7.95	7.59	0.36	(0.144)
Positive reciprocity (2 items, 1-6)	4.61	4.52	4.72	-0.20*	(0.092)
Altruism (donate from EUR 1,000 windfall)	143.53	142.17	144.95	-2.78	(0.884)
Altruism (willingness to share)	9.09	9.21	8.97	0.24	(0.360)
Impulsiveness (1-11)	6.56	6.56	6.57	-0.01	(0.979)
Observations	187	95	92	187	

[47] Power was less of a concern for sales, our other primary outcome variable. This is because we can measure sales at the order level, yielding more precise treatment estimates, whereas labor supply can only be measured at the stylist level.

[48] In the end, stylist recruitment was considerably slower than anticipated by Zalon (averaging around 24 per month). Zalon agreed to leave the experiment in the field longer, so that all applicants with a starting date from October 2016 up to and including July 2017 were part of the experiment. This resulted in approximately the sample size agreed beforehand, namely N=202.

Table A.2: Treatment effect on sales performance by intrinsic motivation, control variables

	(1)	(2)	(3)	(4)
Treated	-17.14***	-12.66**	-11.67***	-12.45***
	(5.686)	(5.003)	(4.404)	(4.302)
Treated × intrinsically motivated	26.70***	22.26***	19.57***	20.67***
	(8.121)	(6.834)	(6.066)	(6.185)
Intrinsically motivated	-20.57***	-19.94***	-17.47***	-19.96***
	(5.747)	(4.877)	(4.017)	(4.103)
Basic controls	Yes	Yes	Yes	Yes
Stylist demographics	No	Yes	Yes	Yes
Customer demographics	No	No	Yes	Yes
Stylist personality	No	No	No	Yes
Adjusted R-squared	0.007	0.016	0.040	0.040
Number of observations	9,524	9,524	9,524	9,524
Number of stylists	187	187	187	187

Standard errors, clustered at stylist level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Outcome variable: normalized order-level sales. Basic controls: randomization stratum, calendar week and potential experience (3rd-degree polynomial). Stylist demographics: age, age squared, gender, abitur, uni, number of kids, expected hours working for Zalon. Customer demographics: age group, country, gender. Stylist personality (median splits of self-reported measures): risk aversion, altruism, impulsiveness (1 item); time preferences, reciprocity (2-item index); conscientiousness (3-item index). F-test treated + treated*intrinsically motivated: (1) p = 0.099, (2) p = 0.041, (3) p = 0.060, (4) p = 0.078.

Table A.3: Treatment effect on sales performance, continuous intrinsic motivation

	(1)	(2)
Treated	-4.43	-74.72***
	(3.946)	(23.115)
Treated × intrinsic motivation index		8.17***
		(2.636)
Intrinsically motivated		-7.29***
		(1.503)
Adjusted R-squared	0.004	0.008
Number of observations	10,090	9,524
Number of stylists	202	187

Note: Standard errors clustered at stylist level in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Outcome variable: normalized order-level sales. Controls: randomisation stratum, calendar week and potential experience (3rd-degree polynomial). Intrinsic motivation (2-item index) is from the baseline survey measure.

Table A.4: Treatment effect on pay satisfaction

	(1)	(2)	(3)	(4)
Treated	-0.94**	-0.83	-0.83**	-0.44
	(0.373)	(0.578)	(0.368)	(0.586)
Treated × intrinsically motivated		-0.19		-0.73
		(0.826)		(0.794)
Intrinsically motivated		0.35		0.62
		(0.594)		(0.577)
Total earnings treatment period (100 EUR)			0.08***	0.09***
			(0.025)	(0.026)
Outcome mean	6.220	6.220	6.220	6.220
Adjusted R-squared	0.061	0.017	0.119	0.084
Number of observations	168	164	168	164

Note: Heteroskedasticity-robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Outcome variable: stylist-level pay satisfaction (1-11). Controls not shown: randomisation stratum, hire month and treatment duration. Intrinsic motivation (2-item index) is a median split from the baseline survey measure.

Table A.5: Correlation value of kept items and intrinsic motivation median split

	(1)
Intrinsically motivated	-1.73
	(1.322)
Adjusted R-squared	0.008
Number of observations	6,394
Number of stylists	186

Note: Standard errors clustered at stylist level in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Outcome variable: order-level value of kept items (EUR). Controls not shown: randomisation stratum, calendar week and potential experience (3rd-degree polynomial). Intrinsic motivation (2-item index) is a median split from the baseline survey measure.

8.3 Additional Material: Formal Model

The agent's choice of quantity and quality:
The first derivatives of the agent's objective function are

$$\beta + \gamma\left(m+q\right) - \left(\frac{\kappa}{2}q^2 - \frac{\eta}{2}\left(\tau - \left(q-q^*\right)^2\right)\right) - \nu n - \frac{1}{2}r\gamma^2\left(2n\sigma_a^2 + \sigma_\varepsilon^2\right) = 0 \tag{2}$$

$$\text{and } \gamma n - n\left(\kappa q + \eta\left(q-q^*\right)\right) = 0. \tag{3}$$

such that (3) becomes

$$q = \frac{\gamma + \eta q^*}{\kappa + \eta}.$$

and we can compute n by rearranging (2) and simplifying to obtain

$$\begin{aligned} n &= \frac{\beta + \gamma\left(m+q\right) - \left(\frac{\kappa}{2}q^2 - \frac{\eta}{2}\left(\tau - \left(q-q^*\right)^2\right)\right) - \frac{1}{2}r\gamma^2\sigma_\varepsilon^2}{\nu + r\gamma^2\sigma_a^2} \\ &= \frac{\beta + \gamma m + \frac{\left(\gamma+\eta q^*\right)^2}{2(\kappa+\eta)} - \eta\frac{q^{*2}-\tau}{2} - \frac{1}{2}r\gamma^2\sigma_\varepsilon^2}{\nu + r\gamma^2\sigma_a^2}. \end{aligned}$$

Proof of Proposition 1:
First note that a shift that keeps the payment per order constant (at prior quality) in the population of agents will imply that

$$E\left[\gamma_0\left(a_i + q_{i0}\right)\right] = \beta + \gamma_1\left(m + E\left[q_{i0}\right]\right) \Leftrightarrow$$

$$\beta = \left(\gamma_0 - \gamma_1\right)\left(m + E\left[\frac{\gamma_0 + \eta q^*}{\kappa + \eta_i}\right]\right).$$

As

$$E\left[\Delta q_i\right] = E\left[\frac{\gamma_1 + \eta_i q^*}{\kappa + \eta_i} - \frac{\gamma_0 + \eta_i q^*}{\kappa + \eta_i}\right] = E\left[\frac{\gamma_1 - \gamma_0}{\kappa + \eta_i}\right]$$

it is clear that there will be a loss in quality and

$$\frac{\partial E\left[\frac{\gamma_1-\gamma_0}{\kappa+\eta_i}\,|\eta_i\right]}{\partial \eta_i} = -\frac{\gamma_1 - \gamma_0}{\left(\kappa + \eta_i\right)^2} > 0$$

such that the loss in quality is the smaller, the more intrinsically motivated an agent is (higher η_i) which shows claims (iii) and (iv).

Now consider the change in quantity

$$\Delta n_i = \frac{1}{\nu}\left(\beta + \gamma_1 a_i + \frac{\left(\gamma_1 + \eta_i q^*\right)^2}{2\left(\kappa + \eta_i\right)} - \eta_i \frac{q^{*2} - \tau}{2} - \frac{1}{2} r_i \gamma_1^2 \sigma_\varepsilon^2 - \left(\gamma_0 a_i + \frac{\left(\gamma_0 + \eta_i q^*\right)^2}{2\left(\kappa + \eta_i\right)} - \eta_i \frac{q^{*2} - \tau}{2} - \frac{1}{2} r_i \gamma_0^2 \sigma_\varepsilon^2\right)\right)$$

$$= \frac{1}{\nu}\left(\beta - \left(\gamma_0 - \gamma_1\right)\left(a_i + \frac{\gamma_1 + \gamma_0 + 2\eta_i q^*}{2\left(\kappa + \eta_i\right)}\right) + \frac{1}{2} r_i \left(\gamma_0^2 - \gamma_1^2\right)\sigma_\varepsilon^2\right)$$

with $\beta = \left(\gamma_0 - \gamma_1\right)\left(m + E\left[\frac{\gamma_0 + \eta_i q^*}{\kappa + \eta_i}\right]\right)$ it becomes

$$\Delta n_i = \frac{1}{\nu}\left(\left(\gamma_0 - \gamma_1\right)\left(m - a_i + E\left[\frac{\gamma_0 + \eta_i q^*}{\kappa + \eta_i}\right] - \frac{\gamma_1 + \gamma_0 + 2\eta_i q^*}{2\left(\kappa + \eta_i\right)}\right) + \frac{1}{2} r_i \left(\gamma_0^2 - \gamma_1^2\right)\sigma_\varepsilon^2\right)$$

Now consider the effect of the treatment on quantity in the population, which is given by

$$E\left[\Delta n_i\right] = \frac{1}{\nu}\left(\left(\gamma_0 - \gamma_1\right)\left(E\left[\frac{\gamma_0 + \eta_i q^*}{\kappa + \eta_i} - \frac{\gamma_1 + \gamma_0 + 2\eta_i q^*}{2\left(\kappa + \eta_i\right)}\right]\right) + \frac{1}{2} E\left[r_i\right]\left(\gamma_0^2 - \gamma_1^2\right)\sigma_\varepsilon^2\right)$$

$$= \frac{1}{\nu}\left(\frac{\left(\gamma_0 - \gamma_1\right)^2}{2} E\left[\frac{1}{\kappa + \eta_i}\right] + \frac{1}{2} E\left[r_i\right]\left(\gamma_0^2 - \gamma_1^2\right)\sigma_\varepsilon^2\right) > 0.$$

Moreover,

$$\frac{\partial E\left[\Delta n_i \left| r_i\right.\right]}{\partial r_i} = \frac{\left(\gamma_0^2 - \gamma_1^2\right)\sigma_\varepsilon^2}{2c_n} > 0.$$

■