The gender gap in illicit earnings*

Max Thon, Oliver Gürtler, Matthias Heinz, Dirk Sliwka**

March 26, 2026

Abstract

Empirical research on gender pay gaps relies primarily on formal payroll data, potentially overlooking a significant component of total economic returns in sectors where illicit income is prevalent, such as the hospitality or the retail industry: We conduct an RCT in a restaurant chain where servers can engage in theft by failing to register self-serving breakfast buffet orders and instead pocketing the customer's payment. Our intervention raises the barriers to this theft by requiring a multiple-employees ordering process. The treatment increases registered buffet orders by 9 percent, an effect driven by those employees most likely to engage in theft prior to the intervention. As physical consumption of core buffet ingredients remains identical between the treatment and control groups, a reduction in employee theft is the only plausible explanation for our findings. The main treatment effect is twice as large for male compared to female servers. In the control group, this illicit income creates a gender gap of 18 percent during breakfast shifts and 10 percent in total monthly income—gaps our intervention effectively eliminates. The treatment induces sorting of workers out of the firm, as male servers are significantly more likely to quit. Ultimately, our study demonstrates how a single compliance measure can fundamentally reshape gender pay gaps.

***JEL Codes**: C93, J71, M52*
***Keywords**: Incentives, gender pay gap, theft, productivity, field experiment*

*We thank numerous colleagues for helpful comments. We further thank our student assistants Marcela Irias, Rosanna Simonis, and Malte Vanderheiden for their great research assistance. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2126/2 – 390838866. The experiment was pre-registered at the registration portal of the American Economic Association with the RCT ID AEARCTR-0014515. The study is approved by the Ethics Committee of the University of Cologne (Reference: 240028MT).

**Max Thon: University of Zurich/University of Cologne, max.thon@business.uzh.ch/m.thon@wiso.uni-koeln.de, Oliver Gürtler: University of Cologne, oliver.guertler@uni-koeln.de, Matthias Heinz: University of Cologne, heinz@wiso.uni-koeln.de, Dirk Sliwka: University of Cologne, sliwka@wiso.uni-koeln.de.

1. Introduction

Quantifying the gender pay gap and identifying its structural drivers is a cornerstone of labor economics. However, empirical research on wage inequality relies almost exclusively on formal payroll data—such as social insurance records, personnel files, and household surveys—potentially overlooking a significant component of total economic returns: illicit income. Despite mounting evidence that such income is both widespread and economically meaningful, we know remarkably little about its magnitude or distribution. Evidence from managerial surveys (Jack L. Hayes International 2023), consultancy reports (Alvarez Marsal Consumer and Retail Group 2024), industry reports (National Retail Federation 2025; National Restaurant Association 2025), and academic research (Chen and Sandino 2012, Pierce et al. 2015, Chan et al. 2021, Brady et al. 2012) suggests that a substantial share of employees in key sectors (in particular Food Services and Retail Trade' sectors—which collectively employ 18–20 percent of the U.S. and EU workforce (U.S. Bureau of Labor Statistics 2025, Eurostat 2024)) — engage in internal theft to increase their total compensation.

Despite this broad-based evidence, empirical studies that quantify illicit wages at the employee level – and thus allow to estimate gender differences – remain rare, largely because employee theft is by its very nature inherently difficult to observe in the field. We address this gap by conducting a Randomized Controlled Trial (RCT) within a large restaurant chain to: (i) quantify illicit wages of female and male employees, (ii) analyze the causal impact of increasing barriers to theft, (iii) evaluate how workers respond to the loss of illicit income (e.g., by quitting the firm or increasing effort to raise legitimate income), and (iv) examine how raising barriers to employee theft ultimately affects gender gaps in performance, illicit wages, and legitimate earnings.

Our study firm is a large, all-day food and beverage chain operating about 80 modern pub-style restaurants in a European country. A restaurant employs, on average, 20 servers and 24 bar/kitchen workers. According to the study firm's top management, employee theft via unregistered delivery of products is a persistent challenge. As is true for most firms in the industry, our firm has implemented a customer ordering process that involves multiple employees to mitigate employee theft: Upon receiving a food or beverage order from a customer, a server enters it into the Point-of-Sales (POS) system, which alerts the bar/kitchen staff. The server then picks up and delivers the prepared items before closing the order in the POS system after payment. This multi-step process, requiring handling by both the server and kitchen/bar staff, limits the opportunity for a single employee to bypass the system and deliver unregistered products, and pocket the customer payment.[1]

A notable exception to this standard order protocol in our firm is the breakfast buffet, which includes classic items like eggs, rolls, sausages, fruits, but also salads and vegetables and is a core product in its offerings. Here, customers serve themselves. This lack of involvement of bar/kitchen staff provides servers a direct opportunity to steal: upon receiving a breakfast buffet order, a server

[1]This includes keeping the cash from a sale or, for card payments, recording the amount as a "tip" instead of a "sale."

can omit entering the buffet properly into the POS system and keep the customer's payment.

As a first step, we show that prior to the intervention female servers registered approximately 9 percent more breakfast buffets than their male counterparts, while sales of other breakfast and lunch offers do not appear to differ. Importantly, this continues to hold when we condition on the sales of hot drinks and other breakfast offerings (which prior to the intervention required registering in the POS). At the same time, male servers register far more unusually high tips—payments exceeding 20 percent of the bill—which is a common way to disguise customer payments for unregistered buffet orders. This appears to be quite unlikely to be driven by truly higher tips given women's comparative advantage in generating tipping income (see, e.g., Chandar et al. 2024).

To capture how much each server deviates from expected buffet sales, we construct a simple, intuitive theft proxy. We predict a server's buffet sales based on their sales of hot drinks and other breakfast items—products that must be entered into the POS and therefore provide a reliable benchmark for actual customer traffic. We then estimate employee fixed effects from this regression. Servers with unusually low buffet sales, relative to what their other sales would predict, receive more negative fixed effects. Multiplying these fixed effects by –1 yields a theft proxy in which higher values correspond to greater suspected theft. Using this measure, we show that servers in the top quartile of the proxy—those whose buffet sales are most unusually low—are 24 percent more likely to be male.

Our intervention aims to mitigate the described theft by involving multiple employees in the breakfast ordering process and making a registered buffet order visually salient. We exploit the fact that the buffet offer includes free juice. While in the control group access to this juice remains self-serve, in randomly selected treatment restaurants, the juice is delivered to customers in a glass from the bar; specifically, to deliver the juice the server has to enter the order in the POS system, the bartender prepares the juice, and the server delivers it to the customer. In the treatment group, it is no longer possible for an individual server to serve the full offer without formal registering the order in the POS. We analyze the treatment effects across 38 restaurants with more than 1,000 servers utilizing fine-grained data from more than 20 million customer orders.

Our intervention increases the total number of registered breakfast buffets by 8.7 percent in the treated restaurants compared to the control group. Importantly, we find that the treatment effects are more than twice as large for male (12.5 percent) than for female servers (5.9 percent) supporting the claim that the gender-based performance gap observed prior to the RCT—and in the control group—is largely driven by differences in theft behavior. We use these results to estimate a gender gap in illicit income which amounts to 17.8 percent during the breakfast shift and 9.8 percent in total monthly income— gaps which our intervention effectively eliminated.

To rule out alternative explanations for the observed patterns, we conduct multiple robustness tests. We, for instance, find no evidence that physical consumption of buffet ingredients increases in the treated stores along with the higher revenues. Similarly, we find no evidence that revenues for other products differ between the two groups during the RCT. Moreover, the treatment effect is most pronounced among servers who were most likely to engage in theft during the pre-RCT

period as measured by our theft proxy.

The increase in registered breakfast buffets in response to our treatment translates to a substantial estimated average revenue gain for the firm (around 90 Euros per restaurant per day), which is, by definition, exactly offset by a corresponding reduction in illicit income for employees who engage in theft.

We also examine whether and how employees adapt to this loss and find that the treatment triggered an increase in attrition among male servers (but not among male bar keepers and kitchen staff). At the same time, we do not observe higher attrition for female servers.

By focusing on gender gaps in illicit income, our paper complements the literature on formal gender pay gaps. Blau and Kahn (2017) document that while the gender wage gap has narrowed over the last several decades, a significant gap remains: women earn approximately 18 percent less than men, and even after adjusting for differences in human capital, occupation, and industry, an unexplained residual difference of nearly 9 percent persists. Research has identified several key measures to further narrow this gap, including pay transparency mandates (Bennedsen et al. 2022), policies that support the reconciliation of family and career (Olivetti and Petrongolo 2017, Kleven et al. 2019), and the reduction of disproportionate rewards for long, inflexible work hours (Goldin 2014, Mas and Pallais 2017). To the best of our knowledge, our study is the first to demonstrate that there is also a substantial gender gap in 'illicit income' within a firm. Furthermore, we provide causal evidence that compliance-enhancing measures can significantly narrow this gap.

Our results complement the body of research documenting gender gaps in criminal behavior, with women generally less likely than men to engage in serious offenses (see, e.g., Campaniello and Gavrilova 2018). Evidence on gender differences in white-collar crime, however, is more limited. For example, Steffensmeier et al. (2013) show that men account for the vast majority of high-level corporate fraud cases, although they note that this pattern may partly reflect the underrepresentation of women in executive positions. We contribute to this literature by providing evidence of substantial gender differences in theft on the job.

Our paper also contributes to the literature on employee misconduct, which highlights the role of wages (Chen and Sandino 2012), peer effects (Chan et al. 2021, Pierce and Snyder 2012), customer characteristics (Balafoutas et al. 2013), and monitoring (Nagin et al. 2002).[2] Most closely related to our work is Pierce et al. (2015), who use a staggered rollout of an IT monitoring tool in a restaurant chain to show that improved monitoring substantially reduces employee theft and increases sales productivity. Although their design and treatment differ from ours, the magnitude of our RCT effects is comparable, underscoring how pervasive theft is in the restaurant sector. Our contribution is to show that theft is not gender-neutral. We document sizable gender gaps in illicit income and tips, and we demonstrate that tightening theft constraints disproportionately reduces these gaps. In this way, we extend the existing literature by revealing the gendered structure of illicit income and how monitoring reshapes it.

[2]Heron et al. (2007) and Zitzewitz (2006) document "informational theft" in the financial industry, where executives and traders exploit backdated options, stale pricing, or insider information to siphon wealth from investors. In contrast, we focus on the theft of cash by low-paid employees and examine gender gaps in the resulting illicit income.

Finally, our study also contributes to the literature on wage cuts, which demonstrates that reducing an employee's regular monthly wage increases turnover (Coviello et al. 2022, Krueger and Friebel 2022). We complement this research by demonstrating that workers are more likely to quit a firm when their illicit wages—rather than their formal wages—are reduced.

2. Background

2.1. Study firm

Our study firm is a large all-day food and beverage chain that operates pub-style restaurants, providing a full range of items, from morning coffee and all-you-can-eat breakfast buffets to burgers and evening cocktails, with prices set at a mid-range level. The firm operates approximately 80 restaurants across one European country, most of them located in urban centers. A typical establishment generates around 200.000 Euros in monthly sales. Menu composition, supply chain, procurement, pricing, standards for customer interaction and food preparation, HR, and marketing campaigns are all centralized by the firm's headquarter.

Each restaurant is overseen by a manager who holds a minority share of the respective store. A typical establishment employs, on average, around 20 servers and 24 bar and kitchen staff members. Around two-thirds of the workforce in the company is female. Wages are fixed and set at or slightly above the national minimum wage level, which corresponds to 13 Euros per hour, in addition to customer tips (corresponding on average to 13.50 Euros per shift).[3] The firm does not pay any performance-based bonuses.

2.2. Order process

A persistent issue in the restaurant industry is the unregistered delivery of products, specifically when servers bypass the point-of-sales (POS) system (Pierce et al. 2015). While the full scope of employee theft is difficult to quantify precisely, available data estimates the annual expense to U.S. firms at between 3 and 6 billion USD (Garber and Walkup 2004). Consequently, order processes throughout the industry are fundamentally designed to minimize the potential for revenue leakage.

In our study firm (as in most firms in the food and beverage industry), upon receiving a food and beverage order from a customer, the server enters it into their handheld device. This action transfers the order to the bar or kitchen through the POS system. The server picks up the prepared food and beverages from the bar or kitchen and delivers them to the customer. After the customer has paid - roughly half of whom use cash and half use card at the time of our intervention - the server closes out the order in the POS system and hands over the receipt to the customer. Servers typically have multiple orders open simultaneously. The multi-party process acts as an internal control: Since both the server and the kitchen and bar staff must handle the order, the opportunity for a single employee to bypass the system and deliver unregistered products is limited.[4]

[3] Average tip figures are derived from card transaction data and exclude implausible values likely stemming from employee theft (see Section 2.4 for the definition of plausible and implausible tips.)

[4] Please see figures A.5, A.6, and A.7 in Appendix C for graphical illustrations of the workflow.

A notable exception, however, is our firm's all-you-can-eat breakfast buffet, which costs about 13 Euros on weekdays and 16 Euros on the weekend. This breakfast buffet is currently offered in 70 stores, which sell, on average, around 60 breakfast buffets per day.[5] The buffet consists of typical breakfast fare, including eggs, sausages, cheese, rolls, salads, vegetables and juice; however, hot drinks (such as coffee or tea) are excluded and must be ordered separately via the regular POS process. The breakfast buffet is typically physically located in the center of the restaurant and, crucially, customers serve themselves. This self-service requirement precludes the involvement of the bar and kitchen staff in the order process. This deviation from the standard ordering process is the primary focus of our research project.

Interviews with managers and staff conducted prior to the RCT revealed a strong suspicion that it is not uncommon among servers to deliberately omit breakfast buffet orders from the point-of-sale (POS) system as consumption here can occur without a digital record. After the customer pays, the server appropriates the revenue. For cash payments, this manipulation leaves no immediate trace. For card payments, however, the discrepancy must be reconciled within the payment terminal, typically by recording the excess amount as a tip. Our main analysis therefore focuses on buffet transactions, while we provide complementary evidence on other product categories where feasible.

2.3. Financial and legal consequences of theft

Even when the firm has collected evidence for theft, establishing legal proof remains difficult. Management can only prove a case if a customer confirms their original order, allowing for a direct comparison with the POS record. This verification is rarely feasible, as customers are typically anonymous and depart quickly after paying; furthermore, a false allegation creates a reputational risk and an unpleasant situation for all parties involved. Additionally, stringent labor laws complicate contract termination even when theft is confirmed. These combined challenges mean that employees face very little risk of formal consequences. In line with Becker (1968) and the empirical findings of Jacob and Levitt (2003), this lack of effective sanctioning severely limits the expected material cost of misconduct. Since servers earn minimum wage, the net gain from stolen money—which is tax-free and exempt from social insurance or pension contributions—can be substantial. This provides a strong incentive for employees to supplement their income through theft.

The financial consequences of theft by servers for the firm are severe because unrecorded revenue has a direct, one-to-one impact on net income. Apart from eroding revenues, the systematic underreporting of sales also promotes a culture of normalized theft and non-compliance and creates a significant legal risk. If an employee steals, the firm does not only lose revenue, but it remains responsible for the unpaid taxes in the legal perspective of the tax authorities. Thus, if employee theft is confirmed, the firm faces the risk of being forced to ex post pay back taxes on the stolen amount and associated fines for tax evasion, necessitating an ex post revision of the financial statements and exposing the firm to further regulatory penalties. The firm recently faced this risk

[5]The remaining stores do not offer the breakfast buffet, primarily due to limited physical space within the restaurant.

firsthand: tax authorities issued inquiries after finding discrepancies between sales and purchasing records, placing the firm under suspicion of tax evasion.

2.4. Indicators of employee theft

As a first step we study gender differences in suspected theft. To do this, we use granular transaction-level data from the firm's POS system for the pre-RCT period (from April 2024) from which we constructed different indicators of employee theft.

Buffet-Hot drink ratio. Hot drinks—such as coffee and tea—are not included in the price of the breakfast buffet and must be ordered separately via the regular POS process, which maintains the multi-party control involving both servers and barkeepers. Hot drinks are a complementary good to the breakfast buffet, as virtually all customers consume the buffet and a hot drink together. Using our fine-grained data, we can measure the number of registered breakfast buffets and the corresponding number of hot drinks ordered by the customers located at the same table by a single server. The core logic of the measure is that systematically lower values for some servers indicate that the server is more likely not to have registered buffet orders in the POS.

Buffet shortfall. Extending this idea further, we construct another theft proxy by capturing how much each server deviates from expected buffet sales. We predict a server's buffet sales based on their sales of hot drinks and other breakfast items—products that must be entered into the POS and therefore provide a reliable benchmark for actual customer traffic. We then estimate employee fixed effects from this regression. Servers with unusually low buffet sales, relative to what their other sales would predict, receive more negative fixed effects.[6] To be precise, we estimate the following regression:

$$BreakfastBuffets_{sith} = \beta_1 \times HotDrinks_{sith} + \beta_2 \times RevenueOtherBreakfast_{sith} + \tau_{shw} + \lambda_i + \epsilon_{sith}, \quad (1)$$

where s indexes restaurants, i servers, t days (or shifts), w weekdays, and h hours. This regression estimates the number of breakfast buffets sold per hour conditional on the number of hot drinks sold as well as the revenue generated from other breakfast products, including time controls (restaurant $\times$ weekday $\times$ hour fixed effects), and most importantly a server fixed effect. We save the estimated server fixed effects λ_i, which capture how much each server systematically overbooks (in case $\lambda_i > 0$) or underbooks (in case $\lambda_i < 0$) breakfast buffets conditional on the number of hot drinks sold and potential substitutes. Multiplying these fixed effects by –1 yields our buffet shortfall measure.

Buffet cancellations. A further method of theft involves initially entering the buffet order into the POS system—displaying it to the customer to create the appearance of proper registration—followed by cancellation immediately after payment is completed. This "immediate cancellation" strategy allows servers to mimic compliant behavior ex ante while removing the transaction ex post. This function is available in the POS to allow servers to correct for mistakes in case

[6] Similar approaches have been used, for example, by Mas and Moretti (2009) or in a crime context by Goncalves and Mello (2026).

they enter the wrong number or product when taking an order. However, unusually many of these immediate cancellations are interpreted as a sign of suspicious behavior and thus equally used as an indicator of theft.

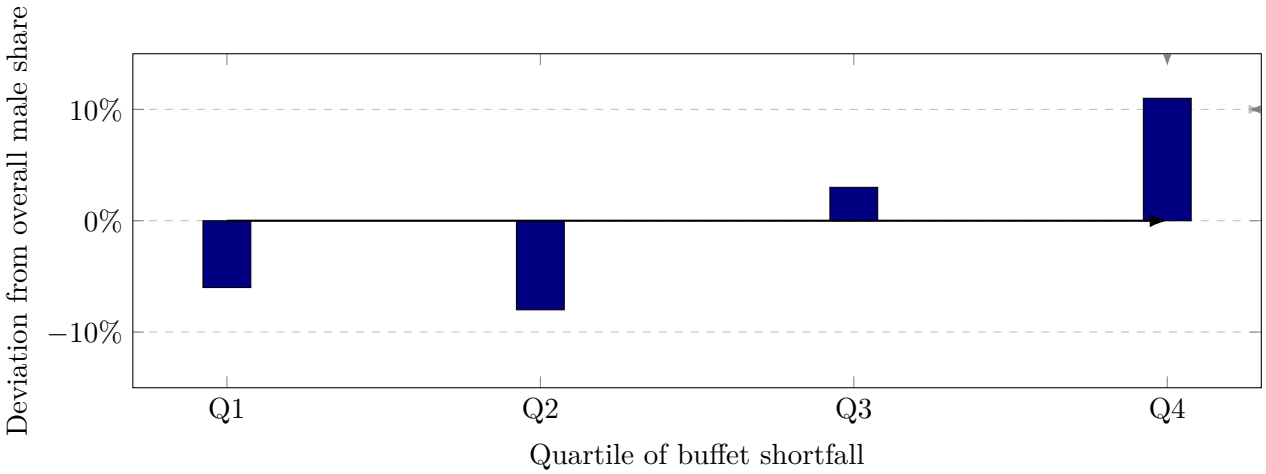
Implausibly large tips. In Appendix A, we plot the distribution of tips as a percentage of the total payment for all card-based transactions in the pre-RCT period and shows that most tips fall into a range consistent with the prevailing social norm of tipping approximately 10 percent. While of course unusually large tips sometimes occur, the firm itself considers systematically registered high tips exceeding 20 percent of the total bill as 'implausible tips' indicating that the server did not register ordered products in the POS to generate own income. It is important to note that we can only identify tips in case guests pay by card. For cash payments, the server does not enter the amount given by the guests into the system.

2.5. Baseline gender gaps in productivity, earnings, and indicators of employee theft

Before outlining our experimental design, we examine baseline gender differences in productivity, earnings, and the indicators of employee theft described in the previous section. Focusing on the pre-intervention period (April–October 2024), we first analyze registrations of the breakfast buffet—the primary outcome variable of our study. While stores averaged 61 buffet sales per day, we find significant gender heterogeneity in the POS data: regressing the number of registered buffets on a gender dummy reveals that women register an average of 1.967 more buffets (+9.3) per day than men (Table 1, Column 1). The firm's top management conventionally interpreted this gap as a difference in productivity; however, an alternative interpretation is that men engage in higher rates of theft by failing to register breakfast buffet orders. Our data provide strong support for this latter interpretation. First, as Column 2 shows, female servers show a significantly lower budget shortfall (that is a lower propensity to fall short of predicted buffet sales) compared to male servers. This significant difference is driven by a gender skew at the upper tail of the buffet shortfall distribution: men are approximately 11 percent overrepresented in the top quartile relative to their overall share in the sample (see Figure 1). Consistent with this finding, we find that the average ratio of breakfast buffets to hot drinks is 4 percent lower for male servers (Table 1, Column 3). Further analysis of our two additional theft indicators—instant cancellations (Column 4) and sum of implausibly large tips (Column 5) — corroborate this pattern. In particular the latter finding appears notable as other studies have documented women's comparative advantages in generating tipping income (see, e.g., Chandar et al. 2024).

Taken together, this baseline evidence supports the view that there may be substantial gender gaps in illicit income due to stronger levels of theft among male employees. In a next step, we provide causal evidence for this claim using an RCT implementing barriers to illicit behavior to study gender differences in employee level outcomes.

Figure 1: Deviation of male share from overall mean by quartile of buffet shortfall



Notes: The figure shows the deviation of the male share in each quartile of the buffet shortfall from the overall male share in the sample. Positive values indicate overrepresentation of male employees relative to the full sample, while negative values indicate underrepresentation.

Table 1: Gender differences in productivity and theft proxies

	Productivity	*Theft indicators*			
	No. Breakfast buffets	Buffet shortfall	Buffet-Hot Drink-Ratio	Cancelations buffet	Sum tips $> 20\%$
	(1)	(2)	(3)	(4)	(5)
Female	1.967**	-0.300**	0.025**	-0.105	-0.701**
	(0.736)	(0.124)	(0.012)	(0.144)	(0.290)
Mean males	21.028	0.190	0.602	1.698	1.856
Observations/Workers	899	899	899	740	899
No. of Clusters	38	38	38	32	38
Controls	No	No	No	Yes	Yes

Notes: This table reports the difference between female and male servers on the respective outcome variables. The unit of observation is at the employee (server) level. All variables are aggregated to the employee level over the pre-RCT period. All specifications include store and a dummy variable indicating whether the server works full time. Standard errors are clustered at the store level and reported in parentheses. The estimated coefficients can be interpreted as marginal effects of being female. Columns (4) and (5) include additional controls. In particular, Column 4 controls for Other Breakfast Revenue and captures revenue from breakfast items other than the buffet (i.e., substitute products). Column 5 controls for the average card share of payments of all transactions. Instant cancellations captures immediately cancelled breakfast buffets. The Buffet-Coffee Ratio is defined as the ratio of breakfast buffets to hot beverages (in quantities). The Buffet shortfall corresponds to the employee-specific fixed effect capturing systematic underreporting behavior, as described in Section 2.4. The first row of additional statistics reports the mean of the dependent variable for male employees in the estimation sample. The number of observations and the number of clusters are also reported. $^{*} < 0.1$, $^{**} < 0.05$, $^{***} < 0.01$

3. The RCT

3.1. Design of our treatment

Our intervention is designed to mitigate employee theft stemming from unregistered breakfast buffet orders by making it substantially harder to serve buffets without registering them in the POS. The intervention leverages the fact that orange juice is included in the buffet and while the juice remains self-serve in the control group, in the treatment group the juice is delivered to customers in a glass from the bar. Specifically, upon a customer placing a breakfast order, the server enters it into the POS system. This action informs the bartender to prepare a juice, which the server then picks up and serves. In contrast, the ordering process remains unchanged in the control group, where customers typically serve themselves juice from the buffet area without requiring a bartender's involvement or a POS entry.

The juice delivered from the bar in the treatment restaurants comes in a distinctive glass purchased specifically for the intervention. This glassware differs from the restaurant's standard collection in its specific form.[7]

3.2. RCT timing

The implementation of the RCT required approximately 200 custom-made glasses for each treatment store, which had to be manufactured by a supplier and delivered to geographically dispersed restaurants. These production and logistical constraints shaped our experimental design in two ways. First, the supplier's limited production capacity restricted the number of restaurants that could receive the treatment at a given point in time. Consequently, the CEO contacted the managers of all 70 restaurants offering the breakfast buffet to ask—without mentioning the RCT—if they were interested in "soon" introducing a new process for buffet orders. Management from 36 restaurants expressed interest and were enrolled in the RCT. Within this sample, the treatment and control groups were randomly assigned at the restaurant level.[8]

The custom-made glasses were then delivered to the 19 randomly selected treatment stores in two waves: around half of the restaurants received the glasses on September 17th or the following few days; the other half received the glasses on October 7th, 2024, or a few days earlier.[9] Restaurants implemented the treatment immediately upon receiving the glasses. The treatment used a neutral framing; workers were simply informed of a change in the buffet process without further details.[10]

[7] The glass incorporates the firm's logo integrated into the glass itself and has a capacity of 0.3 liters but refills were no longer possible, whereas the standard restaurant glassware used for juice in the control group and before the treatment is 0.2 liters and allowed an unlimited refill.

[8] A similar method was, for instance, used by Bloom et al. (2015), in which all employees were asked whether they were interested in working from home, and the treatment group was selected from those who expressed interest.

[9] The delivery timeline was quasi-random, as the distribution was based purely on logistical factors: The supplier delivered glasses to our study firm and other clients during the same logistics trips. Because the supplier lacked internal store data, the delivery order was independent of restaurant characteristics.

[10] According to management, workers viewed the intervention as a measure to improve hygiene, reduce overcrowding, and lower costs by eliminating the unlimited offering. Consistent with the argument that employees did not perceive this as a theft-prevention measure, we found no significant change in perceived trust in management in our employee survey. See Appendix B.

The RCT ran from September 17th to November 3rd, 2024, which serves as the focus of our main analyses.[11]

3.3. Randomization and descriptive statistics

Table 2 gives store-level descriptive statistics for all restaurants (Column 1) and for the control and treatment restaurants (Column 2 and 3) in the pre-RCT period (starting on April 1, 2024).[12] The table also shows standardized mean differences (see Imbens and Rubin 2015) between the control and treatment and p-values from a regression of treatment-group dummies on the outcome variable, including time fixed-effects. The 'full day' variable summarizes data for the entire day, while the 'morning only' variables sum data from 9:00 am to 12:00 pm (Monday to Saturday) and 9:00 am to 1:00 pm (Sunday) as this is the period where the buffet is offered. We provide statistics for registered breakfast buffet orders (our key outcome variables), revenues (overall, and separately for drinks and other food), number of hot coffees and hot drinks, and the average number of servers, bar, and kitchen workers by shift. Similarly, in the last section of the table, based on purchasing data, we show descriptive statistics for three key ingredients used exclusively for the breakfast buffet: bread rolls, sausages and salmon ordered by and delivered to the restaurants (measured in the table in packages). As none of the standardized mean differences exceed 0.25 and none of the coefficients in Column 6 are statistically significant, we conclude that the randomization was successful and that the treatment and control groups are balanced.

3.4. Employee surveys

To measure how the intervention affected the behavior and perceptions of servers and managers, we conducted an employee survey in December 2024 after the RCT (the responses rate is approximately 28 percent). To further understand the professional trajectories of former servers—including their subsequent employers and current jobs—we conducted exit interviews in late 2025. These interviews targeted 30 randomly selected employees who left the firm in the 12 months following the RCT (93 percent response rate).

3.5. Preregistration, ethics approval, statistical power

We registered the intervention in the AEA RCT Registry and obtained approval from the Institutional Review Board of the Faculty of Management, Economics, and Social Sciences at the University of Cologne. Approval was granted without restrictions. We conducted power calculations to ensure our sample is sufficiently large to detect meaningful treatment effects. We determined a minimum detectable effect of 10 percent at a power level of 80 percent. With a sample encompassing 38 restaurants, approximately 1,000 servers, and a dataset of over XY million monthly customer orders, the study is well-powered to identify statistically significant shifts in sales behavior.

[11] Subsequently, from November 4th, 2024, onwards, the treatment was slightly modified due to an additional hot drink offer attached to a buffet order; finally, starting on January 30, 2025, it was rolled out in a staggered fashion to all remaining stores, including those that did not initially express interest in implementing a new ordering process.

[12] In addition to Table 2, we present in Appendix C additional separate balance tables in which we group the stores into an early (see Table A.13) and late cohort (see Table A.14). We similarly conclude that there are no systematic differences with respect to treatment timing.

Table 2: Balance table

	All	Control	Treatment	Absolute Δ	SMD Δ	p-value $\beta_T^{FE} = 0$
Full day						
Revenue	7024.69 (3107.54)	6996.85 (3163.32)	7050.03 (3056.09)	53.18 (664.21)	0.017	0.935
Morning shift only						
Revenue	1785.75 (915.44)	1758.57 (965.01)	1810.50 (867.24)	51.93 (193.48)	0.057	0.757
Quantity buffets	63.05 (36.77)	61.04 (38.53)	64.88 (35.01)	3.84 (7.88)	0.104	0.600
Revenue drinks	650.53 (320.04)	645.48 (338.43)	655.12 (302.30)	9.64 (70.85)	0.030	0.868
Revenue other food	172.74 (114.26)	173.54 (118.72)	172.01 (110.06)	-1.53 (19.35)	-0.013	0.945
No. hot coffee	99.35 (44.71)	101.40 (47.81)	97.48 (41.61)	-3.92 (10.31)	-0.087	0.738
No. hot drinks	107.31 (47.93)	109.59 (51.16)	105.25 (44.71)	-4.34 (11.04)	-0.090	0.734
No waiters by shift	3.59 (1.62)	3.69 (1.74)	3.50 (1.49)	-0.19 (0.35)	-0.118	0.604
No. bar/kitchen workers by shift	4.75 (1.89)	4.75 (1.90)	4.76 (1.89)	0.01 (0.40)	0.005	0.959
Observations	6,826	3,253	3,573			
Input data						
Bread rolls	4.95 (2.77)	4.85 (2.58)	5.04 (2.92)	0.20 (0.64)	0.071	0.749
Sausages	0.85 (0.79)	0.75 (0.74)	0.93 (0.83)	0.18 (0.13)	0.223	0.189
Salmon	0.07 (0.26)	0.11 (0.33)	0.03 (0.18)	-0.08 (0.07)	-0.306	0.228
Observations	1,220	557	663			

Notes: The table provides a summary of the stores' pre-RCT POS, HR, and purchasing data; the sample means comprise daily data for all control observations from April 1, 2024, to October 7, 2024, for both the treatment and the control group. The 'full day' variable summarizes data for the entire day, while the 'morning only' variables sum data from 9:00 to 12:00 (Monday to Saturday) and 9:00 to 13:00 (Sunday). The input data present weekly orderings from the firm's delivery platform (in packages). The table provides the mean for the overall sample (Column 1), the control group (Column 2), the treatment group (Column 3), and the difference between the latter two (Column 4). The last two columns present standardized mean differences (see Imbens and Rubin 2015) and p-values from a regression of treatment-group dummies on the outcome variable, including time fixed-effects.

4. Results of the RCT

4.1. Empirical strategy

We conduct our main analyses at the worker level.[13] For all of the following analyses, we estimate the following equation:

$$y_{it} = \beta D_{it} + \alpha_i + \gamma_t + \epsilon_{it}. \tag{2}$$

where y_{it} represents the respective outcome variable for store/server i on day t, D_{it} is a treatment indicator, α_i captures store/server fixed-effects, γ_t denotes time fixed-effects, and ϵ_{it} is the error term. As the treatment was rolled out in two waves we estimate treatment effects with the imputation estimator of Borusyak et al. (2024), which implements this specification while avoiding the biases of standard two-way fixed effects estimators under staggered treatment timing.[14] This approach is particularly appropriate in our setting because (i) approximately half of the units are never treated, providing a stable control group throughout the sample window, and (ii) outcomes are subject to common demand shocks, so efficiently estimating the untreated outcome process using rich time effects substantially improves precision (Abadie et al. 2025). As advised by Abadie et al. (2023), we cluster standard errors on the unit of randomization, which is the store level.

4.2. Treatment Effects

4.2.1. Effect of the Treatment on Registered Sales

In a first step, we compare the distributions of the number of registered breakfast buffets for treatment and control store observations. Figure 2 clearly shows that the distribution shifts towards higher values in the treatment group. We can reject the null that the two distributions are equal (p-val < 0.0001 for a two-sample Kolmogorov-Smirnov test). The average number of registered breakfast buffets is 7.2 (corresponding to 11 percent) higher in treatment stores compared to control stores.

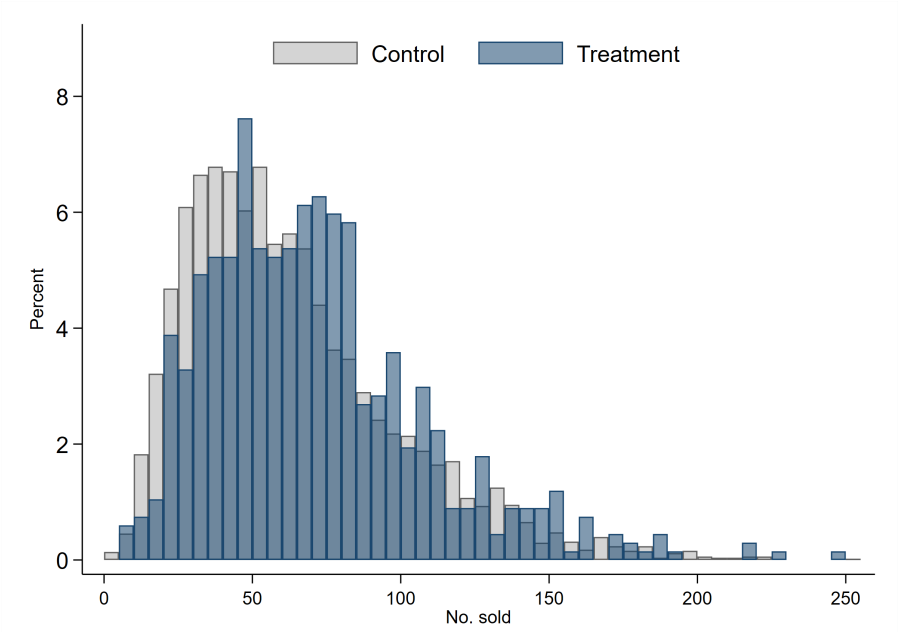
Next, we estimate the average effects of our treatment on our key measure for employee theft: the number of registered breakfast buffets in the POS system. Column 1 of Table 3 provides the result. We find an economically and statistically significant treatment effect as the mean number of registered breakfast buffet orders increases by 1.93 (corresponding to an increase of 8.7 percent) per day per worker when the treatment is implemented.[15]

[13] We also replicate the main analyses on restaurant level in Table A.9 in Appendix A.

[14] In Table A.10 in Appendix A, we also report results from standard two-way fixed effects regressions which lead to similar results.

[15] This treatment effect is equivalent to 0.118 of the standard deviation of the mean number of registered breakfast buffets in the control period.

Figure 2: Distribution of breakfast buffet sales by store per day



Notes: This graph shows a histogram of the number of sold breakfast buffets by store per day. The red bars show the observations from the treatment group and the blue ones from the control group. A two-sample Kolmogorov–Smirnov test checking for equality of distributions yields a two-sided p-value of 0.000.

Table 3: Main treatment effect on registered buffet sales

	No. breakfast buffets		
	All	*Female*	*Male*
	(1)	(2)	(3)
Treatment	1.930***	1.475***	2.893***
	(0.594)	(0.552)	(0.675)
Control mean	24.370	24.799	23.251
p-val Female = Male		0.012	
p-val $H_0 : \tau = 0$	0.001	0.008	0.000
Observations	19,032	19,032	19,032
No. of Clusters	38	38	38

Notes: This table presents treatment effects on sales outcomes at the worker level. The outcome variable in Column 1 is the number of sold breakfast buffets. Column 2 presents the effect on the daily food-drink ratio of the worker, which is the number of breakfast buffets divided by the number of hot drinks sold. The estimates are obtained using the imputation estimator of Borusyak et al. (2024); the respective marginal effects need to be interpreted in terms of marginal increases. The first row of additional statistics show the p-value for the test of the treatment effect being equal to zero. Standard errors clustered on store level are reported in parentheses. *< 0.1, **< 0.05, ***< 0.01

As discussed in Section 2.5, prior to our RCT, female servers had registered approximately 9.3 percent more breakfast buffets than their male counterparts. As laid out in the above this – along with the results on the other theft proxies – already provided an initial indication that male servers engage in more employee theft via unregistered breakfast buffet orders. Our RCT results provide

causal evidence for this gender gap in theft. Columns 2 and 3 of Table 3 show the treatment effects separated by gender. In particular, Column 2 shows that female servers register on average 1.475 more breakfast buffets per day (corresponding to 5.9 percent). Column 3 reports the effect for male servers, which is almost twice as large, with an estimated increase of 2.893 per day (corresponding to 12.54 percent) and this difference in the treatment effects between both genders is significant ($p = 0.012$).

We can use these treatment effects to estimate the magnitude of the gender gap in illicit income prior to the intervention. When men, on average, pocket about $1.418 \cdot 13.38$ euros (the weighted price over the week) more of the buffet revenue, their daily average earnings were higher by about €2.37, or 18 percent of the average hourly wage (assuming an 8-hour morning shift). Extrapolating this to a full day with 16 hours of store operation (assuming there are no theft opportunities during late shifts), this translates into an expected difference in hourly wages of 10 percent.[16]

4.3. Robustness checks and alternative explanations

We argued in the above that the treatment is driven by the increased barriers to employee theft via the modified ordering process in treated restaurants, which shifted otherwise unregistered breakfast buffet orders onto the POS system rather than being appropriated by employees. But of course it is important to test the robustness of this claim and investigate alternative potential explanations which we do in the following.

4.3.1. Buffet inputs

We compare the physical ingredients consumed in the breakfast buffet between the treatment and control groups using restaurant-level purchasing data. Measuring exact daily inputs for the entire buffet is challenging, as many ingredients are used across multiple menu items (e.g., tomatoes for both the buffet and burgers) or can be stored for several weeks (e.g., butter). However, bread rolls, sausages, and salmon serve as useful proxies: they are used exclusively for the breakfast buffet, consumed by nearly all guests, and cannot be stored long-term. As shown in Table 4, neither for the number of rolls nor for sausages and salmon we find significant treatment differences. Hence, we find no evidence that physical consumption is affected by the treatment supporting the view that it is not the number of ordered and consumed buffets that is increased by the treatment but the number of those buffets registered in the POS.

4.3.2. Treatment effects by prior buffet shortfall

In Section 2.4, we estimated an indicator for an individual server's propensity to engage in buffet theft. This residual-based buffet shortfall indicator captures whether servers systematically registered fewer breakfast buffets conditional on complementary product consumption (e.g., hot drinks) and substitutes for the breakfast buffet. Figure 3 presents treatment effects aggregated for workers

[16] See Appendix C for a detailed calculation.

Table 4: Treatment effects on buffet inputs

	Ordered products for buffet		
	Bread rolls	Sausages	Salmon
	(1)	(2)	(3)
Treatment	0.157	0.007	-0.013
	(0.179)	(0.057)	(0.009)
Control mean	4.944	0.832	0.073
p-val $H_0 : \tau = 0$	0.380	0.897	0.143
Observations	1,220	1,220	1,220
No. of Clusters	38	38	38

Notes: This table presents treatment effects on weekly order inputs by restaurants. The outcome variable is the number of ordered packages. Column 1 presents the effect on ordered bread rolls, Column 2 on ordered sausages, and Column 3 on ordered salmon packages. The estimates are obtained using the imputation estimator of Borusyak et al. (2024); the respective marginal effects need to be interpreted in terms of marginal increases. The first row of additional statistics show the p-value for the test of the treatment effect being equal to zero. Standard errors clustered on store level are reported in parentheses. $^{*}< 0.1$, $^{**}< 0.05$, $^{***}< 0.01$

based on the quartiles of the value of their corresponding budget shortfall prior to the intervention.[17] $Q1$ corresponds to the 25 percent of workers with the lowest value of λ_{si} and thus those least likely engaging in employee theft in the pre-RCT period, while $Q4$ captures the 25 percent of workers with the highest value of λ_{si} and thus those most likely to engage in such behavior. $Q2$ and $Q3$ are constructed analogously. As the figure indicates, we find the largest treatment effects for servers belonging to the highest quartile $Q4$, that is the treatment effects are most pronounced for those servers with the highest likelihood of having engaged in not registering breakfast buffet orders prior to the experiment.[18]

As we have shown before (see Figure 1), men are substantially overrepresented in the top quartile of our buffet shortfall theft proxy. In a next step, we investigate gender differences *within* each quartile. Table 5 shows treatment effect estimates by budget shortfall quartiles and gender. While we still see some within quartile gender differences, they are substantially reduced, in line with the view that gender differences in the treatment effects are largely driven by differences in the propensity to steal at the outset.

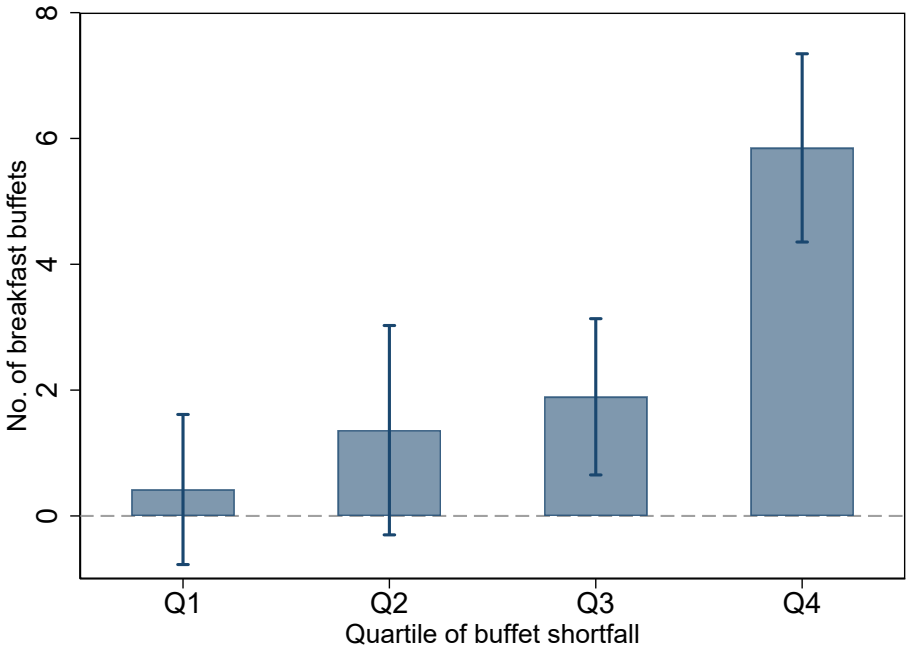
4.3.3. Other theft indicators

In 2.4 we introduced three additional indicators of theft: (i) Because breakfast buffets are typically consumed together with a hot drink, failing to register buffet orders should manifest as an unusually low *buffet–hot-drink ratio*. (ii) Servers may temporarily register a buffet in the POS and cancel it later—after the customer has paid—in order to pocket the payment. We therefore track the number of *buffet cancellations*. (iii) Servers may also record the payment for a buffet as a tip,

[17] See Table A.11 in Appendix A for the corresponding regression table.

[18] Note that this cannot be driven by a regression-to-the-mean effect, as any such effect should also be present in the control group.

Figure 3: Treatment effect by buffet shortfall



Notes: This figure shows a coefficient plot of the impact of the treatment on the number of sold breakfast buffets by worker by quartile of buffet shortfall. For instance, $Q4$ shows the aggregated treatment effects of the 25 percent workers proxied as having the highest likelihood of engaging in illegal behavior due to high buffet shortfall prior to the RCT. $^{*}< 0.1$, $^{**}< 0.05$, $^{***}< 0.01$

Table 5: Main treatment effect by theft proxy and gender

	No. breakfast buffets			
	Q1–Q3		Q4	
	Female (1)	Male (2)	Female (3)	Male (4)
Treatment	1.014*	1.951***	5.036***	6.884***
	(0.563)	(0.712)	(1.001)	(0.452)
Control mean	26.130	25.311	19.583	18.649
p-val $H_0 : \tau_{Male} = \tau_{Female}$	0.105		0.040	
Observations	19,032			
No. of Clusters	38			

Notes: This table shows the impact of the treatment on the number of breakfast buffets sold per worker, by quartile of the buffet shortfall measure. Column 1 ($Q1$–$Q3$) shows the aggregated treatment effects for the 75 percent of workers proxied as having a relatively lower likelihood of engaging in illegal behavior prior to the RCT. Column 4 ($Q4$) presents the treatment effect for the 25 percent of employees with the highest likelihood of engaging in illegal behavior prior to the RCT. The estimation includes server and time fixed effects. The estimates are obtained using the imputation estimator of Borusyak et al. (2024); the respective marginal effects should be interpreted as marginal increases. The first row of additional statistics in both panels shows the p-value for the test of equality of the treatment effect between female and male servers. Standard errors clustered at the store level are reported in parentheses. $^{*}< 0.1$, $^{**}< 0.05$, $^{***}< 0.01$

allowing them to keep the money which results in a higher share of *unusually high tips*.

Table 6 reports the treatment effects on these three indicators again separately for the top quartile of servers with high buffet shortfall and the remaining sample. As the table shows, in the top quartile – i.e., among those servers predicted to be most likely to steal – the treatment significantly reduces the number of *buffet cancellations* (Panel A) by about 23.4 percent and increases the *buffet–hot-drink ratio* (Panel B) by 17.5 percent. However, we do not observed a decrease in the number of *unusually high tips* (Panel C). A potential explanation for the latter result is that servers who had engaged in buffet theft before the intervention now engage in other (legal or illegal) ways to increase their income to compensate for the shortfall of income from buffet theft.

4.3.4. Other robustness checks

The main qualitative results are the same when we use the standard TWFE-Estimator (using time and worker fixed-effects), and when restricting the analysis to the period following the staggered rollout, when all treatment stores were simultaneously treated. Alternative estimation approaches, including Poisson regression models (Correia et al. 2020) and an alternative form of difference-in-difference estimator (Callaway and Sant'Anna 2021), yield qualitatively similar results.[19] Lastly, we conducted a placebo check, by investigating potential treatment effects on revenue from other drinks and food besides the breakfast. We find no significant treatment effects on both outcomes. All of these analyses can be found in Table A.10 in Appendix A.

4.4. Does reducing theft opportunities drive worker attrition?

The approximately 9 percent increase in the total number of registered breakfast buffets translates into an estimated average revenue gain of 90 Euros per five-hour shift in each restaurant (see Appendix C). This revenue gain is, by definition, exactly offset by a corresponding reduction in illicit income for employees who engage in theft. For an individual worker, we estimate that this hourly illicit income loss averages 23 percent of their formal hourly wage (during morning shifts)—if they appropriate 1.93 (ATE) breakfast buffet per shift.

Standard reasoning would imply increased turnover among employees due to the reduction in earnings capacity: Employees who have generated substantial additional income through theft may want to move to different jobs either in the same industry or to other restaurants where theft is still possible.

We thus investigate whether the reduction in "illicit income" induced by the treatment affects server attrition. To do so, we estimate the treatment effect on a binary indicator for whether an employee quit the firm in a given month between the start of the RCT and the end of January 2025,

[19] Estimates based on the Callaway and Sant'Anna (2021) estimator are similar in magnitude but less precise. This pattern is consistent with the low serial correlation observed in our outcome variable. When shocks are largely transitory, regression-imputation estimators efficiently use information from multiple pre-treatment periods to estimate counterfactual trends, while subgroup difference-in-differences estimators rely only on the last pre-treatment period, resulting in lower statistical efficiency (Harmon 2022). In our data, residual autocorrelations decay quickly across lags, indicating weak persistence of shocks.

Table 6: Treatment effect on further indicators of theft

Panel A: Instant cancellations buffet		
	Q1–Q3	Q4
	(1)	(2)
Treatment	-0.114	-0.467***
	(0.093)	(0.174)
Control mean	1.732	2.001
p-val $H_0 : \tau_{Q4} = \tau_{Q1\text{–}Q3}$	0.050	
Observations	16,091	
No. of Clusters	32	
Panel B: Food-Hot Drink-Ratio		
	Q1–Q3	Q4
	(1)	(2)
Treatment	0.009	0.085***
	(0.015)	(0.023)
Control mean	0.643	0.486
p-val $H_0 : \tau_{Q4} = \tau_{Q1\text{–}Q3}$	0.001	
Observations	18,989	
No. of Clusters	38	
Panel C: Sum tips $> 20\%$		
	Q1–Q3	Q4
	(1)	(2)
Treatment	0.307	0.472
	(0.297)	(0.380)
Control mean	1.278	1.182
p-val $H_0 : \tau_{Q4} = \tau_{Q1\text{–}Q3}$	0.513	
Observations	19,032	
No. of Clusters	38	

Notes: This table shows the impact of the treatment on other theft indicators, by quartile of the buffet shortfall measure. Column 1 ($Q1$–$Q3$) shows the aggregated treatment effects for the 75 percent of workers proxied as having a relatively lower likelihood of engaging in illegal behavior prior to the RCT. Column 4 ($Q4$) presents the treatment effect for the 25 percent of employees with the highest likelihood of engaging in illegal behavior prior to the RCT. The estimation includes server and time fixed effects. The estimates are obtained using the imputation estimator of Borusyak et al. (2024); the respective marginal effects should be interpreted as marginal increases. The first row of additional statistics in both panels shows the p-value for the test of equality of the treatment effect between female and male servers. Standard errors clustered on store level are reported in parentheses. $^{*}< 0.1$, $^{**}< 0.05$, $^{***}< 0.01$

as recorded in the firm's personnel files.[20] Panel A of Table 7 presents these results separately by

[20]Our attrition analysis includes the Secondary Intervention Phase (where buffet offers came along with a free hot drink) because: (i) employee turnover typically involves a lag due to stipulated notice periods, (ii) resignations often align with the end of the calendar year, and (iii) the treatment remained virtually identical from the employees' perspective (while the customer interface underwent minor adjustments, the employees' daily work and incentives were functionally invariant, making it highly unlikely that these modifications differentially affected turnover). Our main qualitative findings remain robust when restricting the analysis to the primary RCT period.

gender. We find that the treatment (weakly) increases attrition among male servers (Column 1), but has no discernible effect on females (Column 2). The magnitude is economically substantial: men in the treatment group are 50 percent more likely to quit than those in the control group.

Table 7: Treatment effects on turnover and turnover intention

Panel A: Turnover (Administrative data)			
	All	Female	Male
	(1)	(2)	(3)
Treatment	0.005	-0.012	0.038*
	(0.013)	(0.016)	(0.022)
Control mean	0.069	0.070	0.066
p-val $H_0 : \tau = 0$	0.705	0.461	0.076
Observations	7,289	4,893	2,396
No. of Clusters	38	38	38
Panel B: Standardized turnover intention (Survey data)			
	All	Female	Male
	(4)	(5)	(6)
Treatment	0.249	0.110	0.460*
	(0.152)	(0.200)	(0.235)
Control mean	-0.099	-0.031	-0.209
p-val $H_0 : \tau = 0$	0.104	0.585	0.054
Observations	183	111	72

Notes: This table presents treatment effects on (Panel A) actual turnover and (Panel B) turnover intention at the worker level. In Panel A, the dependent variable is a dummy equal to one if a server quits the firm in a given month, based on administrative personnel records. The specifications include store and time fixed effects. In Panel B, the dependent variable is a standardized index of turnover intention based on survey responses. The index combines two questions: (i) the likelihood of leaving the job within six months (scale 1–7), and (ii) the frequency of thinking about quitting (scale 1–7). The index is standardized to have mean zero and standard deviation one. Estimates can be interpreted as marginal effects. Standard errors are reported in parentheses and clustered at the store level in Panel A and robust in Panel B. *< 0.1, **< 0.05, ***< 0.01

As a placebo check, we also examine the turnover of barkeepers and kitchen staff. These employees operate in the same workplace environment but, unlike servers, have significantly fewer opportunities to steal via unregistered breakfast buffets. We find no significant treatment effect on the attrition of these groups.[21]

Finally, we examine whether these actual quits are mirrored by subjective turnover intentions, using data from our employee survey conducted during the RCT. We surveyed servers on their self-

[21]While barkeepers' workloads increased slightly due to the task of filling juice glasses, this requirement added only a few seconds per glass. Furthermore, our mid-RCT survey reveals no significant differences in reported stress levels across barkeepers, or kitchen staff between the treatment and control groups (see Appendix B). This lack of a treatment effect on stress, combined with the stable attrition rates for these groups, strongly suggests the intervention did not meaningfully increase job disutility or alter the non-pecuniary aspects of their daily work.

reported likelihood of leaving the firm. We report the results in Panel B of Table 7.[22] Consistent with the results on actual turnover, male servers in the treatment group report a 0.46 higher standard deviation in turnover intention than the control group. For female servers, the effect is smaller and statistically insignificant.[23] Taken together, our results demonstrate that male servers in the treatment group exhibit both a higher intention to leave and a significantly higher actual quit rate.

As servers who left the firm primarily moved to comparable jobs in the same sector (see exit interview results in Appendix F), our findings suggest that the intervention creates a negative externality: by driving workers prone to employee theft from our study firm toward competitors, the treatment changes not only the composition of the workforce in our firm, but also increases the share of such workers at competing firms within the local labor market.[24]

5. Conclusion

This paper provides causal evidence that a significant portion of the gender pay gap can be hidden within "illicit income"—informal earnings derived from workplace misconduct that remain invisible in formal payroll data. Through a large-scale RCT, we demonstrate that a single, compliance-enhancing measure significantly increases firm revenues while simultaneously mitigating gender-based income inequality, and inducing potentially beneficial worker sorting.

Our findings show that the pre-intervention "productivity gap," where female servers appeared to outperform their male counterparts in official records, was primarily driven by male servers disproportionately engaging in employee theft. We estimate that this difference in misconduct created a 10 percent gender gap in total monthly income, functioning as an illicit and gendered wage premium for men. By increasing the visual saliency of orders and involving the bar staff, our intervention effectively compressed this illicit income gap, providing a more accurate and equitable measure of true labor market performance.

The elimination of this illicit income led to significant changes in the workforce and its behavior.

[22] To measure employees' intention to quit, we conducted a survey four weeks into the RCT. Servers were asked two questions: (i) "How likely is it that you will quit your job within the next six months?" and (ii) "How often do you think about quitting your job at the restaurant?" Both questions were answered on a 7-point Likert scale, where higher values indicate a stronger intention to quit. We construct a composite measure of quitting intentions by summing both items and standardizing the resulting index to have mean zero and standard deviation one in the sample. Higher values indicate a greater propensity to leave the job.

[23] These survey results also strongly suggest that the observed turnover is voluntary. This is supported by the fact that firing rates remained near zero across both groups. Furthermore, it is unlikely that managers "nudged" servers to quit via assignment to undesirable shifts and restaurant tables (see Pierce et al. 2015), as the treatment did not alter the manager's individual detection probability of specific employees.

[24] While one could imagine a scenario where more productive servers engage in higher rates of theft—with management tacitly permitting it as a performance-based 'bonus'—such a strategy is highly improbable in practice. Tolerating such misconduct is a high-risk strategy, as it can foster a culture of non-compliance (Pierce et al. 2015, Chan et al. 2021) and expose the firm to legal risks, including potential fraud charges. Measuring server productivity precisely is challenging in our setup because standard performance metrics could be correlated with theft. Our preferred productivity measure is the average plausible tip (controlling for gender), as it reflects service quality from the customer's perspective. In our RCT, we find that the treatment effects on registered breakfast buffet orders are the same for workers with above- and below-median average tips in the pre-RCT period (see Appendix A), suggesting that more productive servers do not steal more.

First, we identify a worker selection effect: the treatment caused an increase in attrition among male servers. This suggests that increasing barriers to theft leads to the targeted sorting of dishonest workers out of the firm. Because these departing workers frequently remain within the hospitality sector, our findings suggest that such interventions may create a negative externality, effectively shifting the burden of misconduct onto competitors.

Ultimately, our intervention highlights that quantifying the gender pay gap in industries prone to employee theft—such as restaurant and retail—requires looking beyond formal social insurance records and personnel files. Our study underscores that enhancing structural compliance is not only a matter of financial oversight but a critical step toward achieving true gender equity and meritocracy in the labor market.

While our study provides causal evidence of a gender gap in illicit income, the underlying psychological or sociological mechanisms driving the gender difference in illicit behavior remain beyond our empirical scope. Explanations why male servers exploit systemic vulnerabilities at higher rates than their female colleagues could stem from gender differences in overconfidence, risk preferences, peer effects within female and male groups, image concerns, or preferences for (dis)honesty. We leave the task of isolating these specific behavioral drivers to future research.

References

Abadie, A., J. Angrist, B. Frandsen, and J.-S. Pischke (2025). Harvesting Differences-in-Differences and Event-Study Evidence. *National Bureau of Economic Research*. Working paper no.34550.

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics 138*(1), 1–35.

Alvarez Marsal Consumer and Retail Group (2024). Retail's trojan horse: Tackling the complexities of internal shrink. Accessed: 2026-03-01.

Balafoutas, L., A. Beck, R. Kerschbamer, and M. Sutter (2013). What drives taxi drivers? A field experiment on fraud in a market for credence goods. *Review of Economic Studies 80*(3), 876–891.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy 76*(2), 169–217.

Bennedsen, M., E. Simintzi, M. Tsoutsoura, and D. Wolfenzon (2022). Do firms respond to gender pay gap transparency? *The Journal of Finance 77*(4), 2051–2091.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*(3), 789–865.

Borusyak, K., X. Jaravel, and J. Spiess (2024). Revisiting Event-Study Designs: Robust and Efficient Estimation. *The Review of Economic Studies 91*(6), 3253–3285.

Brady, M. K., C. M. Voorhees, and M. J. Brusco (2012). Service Sweethearting: Its Antecedents and Customer Consequences. *Journal of Marketing 76*(2), 81–98.

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Campaniello, N. and E. Gavrilova (2018). Uncovering the gender participation gap in crime. *European Economic Review 109*, 289–304.

Chan, T. Y., Y. Chen, L. Pierce, and D. Snow (2021). The influence of peers in worker misconduct: Evidence from restaurant theft. *Manufacturing & Service Operations Management 23*(4), 952–973.

Chandar, B., U. Gneezy, J. A. List, and I. Muir (2024). Social preferences in the field: Evidence from a nationwide tipping study.

Chen, C. X. and T. Sandino (2012). Can Wages Buy Honesty? The Relationship Between Relative Wages and Employee Theft. *Journal of Accounting Research 50*(4), 967–1000.

Correia, S., P. Guimarães, and T. Zylkin (2020). Fast Poisson estimation with high-dimensional fixed effects. *The Stata Journal 20*(1), 95–115.

Coviello, D., E. Deserranno, and N. Persico (2022). Minimum wage and individual worker productivity: Evidence from a large us retailer. *Journal of Political Economy 130*(9), 2315–2360.

Eurostat (2024). Employment rate and sectoral employment statistics, eu, 2024. Accessed: 2026-03-01.

Garber, A. and C. Walkup (2004). Theft protection plan key to locking in profits. *Nation's Restaurant News 38*(21), 92–98.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American economic review 104*(4), 1091–1119.

Goncalves, F. and S. Mello (2026). Police Discretion and Public Safety. *Journal of Political Economy Microeconomics*. Working Paper no. 31678.

Harmon, N. A. (2022). Difference-in-Differences and Efficient Estimation of Treatment Effects. *Work*, 241–264.

Heron, R. A., E. Lie, and T. Perry (2007). On the use (and abuse) of stock option grants. *Financial Analysts Journal 63*(3), 17–27.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jack L. Hayes International, I. (2023). 35th annual retail theft survey: Shoplifter and dishonest employee apprehensions and dollar recoveries soared in 2022 as 81% of surveyed participants reported an increase in shrink. Accessed: 2026-03-01.

Jacob, B. A. and S. D. Levitt (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics 118*(3), 843–877.

Kleven, H., C. Landais, J. Posch, A. Steinhauer, and J. Zweimüller (2019). Child penalties across countries: Evidence and explanations. In *Aea papers and proceedings*, Volume 109, pp. 122–126. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Krueger, M. and G. Friebel (2022). A pay change and its long-term consequences. *Journal of Labor Economics 40*(3), 543–572.

Mas, A. and E. Moretti (2009). Peers at work. *American Economic Review 99*(1), 112–45.

Mas, A. and A. Pallais (2017). Valuing alternative work arrangements. *American Economic Review 107*(12), 3722–3759.

Nagin, D. S., J. B. Rebitzer, S. Sanders, and L. J. Taylor (2002). Monitoring, motivation, and management: The determinants of opportunistic behavior in a field experiment. *American Economic Review 92*(4), 850–873.

Olivetti, C. and B. Petrongolo (2017). The economic consequences of family policies: lessons from a century of legislation in high-income countries. *Journal of Economic Perspectives 31*(1), 205–230.

Pierce, L., D. C. Snow, and A. Snyder (2015). Cleaning house: The impact of information technology monitoring on employee theft and productivity. *Management Science 61*(10), 2299–2319.

Pierce, L. and J. A. Snyder (2012). Discretion and manipulation by experts: Evidence from a vehicle emissions policy change. *The B.E. Journal of Economic Analysis & Policy 13*(3), 1–32.

Steffensmeier, D. J., J. Schwartz, and M. Roche (2013). Gender and twenty-first-century corporate crime: Female involvement and the gender gap in Enron-era corporate frauds. *American Sociological Review 78*(3), 448–476.

U.S. Bureau of Labor Statistics (2025). Employment by major industry sector: Percent distribution of total employment, 2024. Accessed: 2026-03-01.

Zitzewitz, E. (2006). How widespread was late trading in mutual funds? *American Economic Review 96*(2), 284–289.