The gender gap in illicit income*

(Preliminary and early stage work - Please do not circulate without permission)

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

September 6, 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 industry. We conduct an RCT in a restaurant chain where servers can engage in theft by failing to register self-served 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 servers most likely to engage in theft prior to the intervention. As physical consumption of core buffet ingredients does not differ between the treatment and control groups, a reduction in employee theft is the only plausible explanation for our findings. The treatment effect is twice as large for male compared to female servers. Servers' illicit earnings from unregistered buffet orders create a conservatively estimated 10 percent gender gap in total monthly income in the control group—a gap that is eliminated in treated restaurants. Male servers adapt to the loss of illicit income with higher attrition, while those remaining who lost the most income shift to stealing directly from customers, generating a 2 percentage point gender gap in illegal income. This shift in theft from the firm to customers introduces reputational risks for the chain, illustrating how a compliance measure can backfire.

***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/MPI for Behavioral Economics, 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 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 within key sectors of the economy, we know remarkably little about its magnitude and distribution. Evidence from managerial surveys (Jack L. Hayes International 2023), consultancy reports (Alvarez Marsal Consumer and Retail Group 2024), industry reports (NRF 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 settings where staff directly handle customer payments—most notably Food Services and Retail Trade, which collectively employ around 20 percent of the U.S. as well as the EU workforce (U.S. Bureau of Labor Statistics 2025, Eurostat 2024)—engage in internal theft to increase their total compensation.

Empirical studies that quantify illicit income at the employee level—and thus allow for estimating gender differences—remain rare, largely because employee theft is, by its very nature, 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 the magnitude of employee theft and illicit income of female and male servers in the firm, (ii) estimate the causal impact of increasing barriers to theft on misconduct, performance and gender gaps, and (iii) analyze how servers respond to the loss of illicit income (e.g., by quitting the firm, substituting other forms of legal or illicit income).

Our study firm is a large, all-day food and beverage chain operating about 70 modern pub-style restaurants in a European country, all of which offer an all-you-can-eat breakfast buffet. On average, a restaurant employs 20 servers and 24 bar/kitchen workers, earning the statutory minimum wage plus a fixed supplement for shift supervisors—a pay scheme that leaves virtually no room for a gender difference in formal pay.

As is common across the hospitality industry (Pierce et al. 2015), employee theft is—according to the firm's top management—a persistent challenge. Because theft often occurs when an employee delivers a product to a customer without recording the sale and pockets the payment instead, the firm has implemented a customer ordering process involving multiple employees to mitigate the risk: Upon receiving a food or beverage order from a customer, a server enters it into the Point-of-Sale (POS) system, which alerts the bar/kitchen staff. The server then picks up and delivers the prepared items to the customer and later collects payment. This 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's payment.

A notable exception to this standard order protocol in our firm is the breakfast buffet, which includes classic items like eggs, rolls, and fruits, and is a core product offering. Here, customers serve themselves. This lack of involvement of bar/kitchen staff provides servers with a direct opportunity to steal: upon receiving a breakfast buffet order, a server can simply omit entering the

buffet into the POS system and keep the customer's payment—either by pocketing the cash or, for card payments, recording the amount as a tip instead of a sale.

For the firm, detecting theft at the buffet—as with other methods of theft—is difficult, since establishing legal proof requires a customer to confirm their original order, and customers are typically anonymous and depart quickly. Even when theft is detected, stringent labor laws make formal sanctioning legally restrictive. The setting thus reproduces the central trade-off of the canonical economic model of crime (Becker 1968): low expected costs of misconduct against substantial, untaxed gains on top of a near-minimum wage.

Looking at the pre-intervention period, we find several strong indications that servers indeed respond to these incentives for misconduct. In particular, we design a proxy for theft that captures how much each server deviates from expected buffet sales, based on their sales of hot beverages—a complementary good that virtually all buffet customers also order, but which must go through the standard multi-employee ordering process. This measure reveals that a significant share of servers registered fewer breakfast buffets than expected given their hot beverage sales, consistent with pocketing the corresponding payments before our intervention. Turning to whether this misconduct varies systematically across genders, and consistent with a broader literature documenting that men are more prone to illicit behavior (e.g., Steffensmeier and Allan 1996, Campaniello and Gavrilova 2018), servers in the top quartile of this theft proxy are about 10 percent more likely to be male than the average server (35 versus 32 percent). Indeed, male servers registered approximately 9 percent fewer breakfast buffets than their female counterparts before the RCT—a gap the firm's top management initially attributed to women being more productive at selling the buffet.

Our intervention aims to mitigate misconduct at the breakfast buffet by involving multiple employees in the ordering process. We exploit the fact that the buffet offer includes juice. While in the control group access to this juice remains self-serve, randomly selected treatment restaurants shift the juice to table service: the server must enter the order into the POS system, the bar staff prepares the juice in a glass, and the server brings it to the customer. Thus, in the treatment group, it is no longer possible for an individual server to deliver the entire buffet without formally registering the order in the POS. We analyze the treatment effects in an RCT involving 38 restaurants with more than 1,000 servers, utilizing fine-grained transaction-level data on every customer order.

Turning to the results, treated restaurants register 9 percent more breakfast buffets than control restaurants, with the effect being around twice as large for male servers (12.5 percent) than for female servers (6 percent). These results translate to a gender gap in illicit income in the control group amounting to about 19 percent during the morning shift and 10 percent in total monthly income stemming from unregistered breakfast buffet orders—a gap that is eliminated in the treatment group.

The treatment effect is largest among the predominantly male group of servers classified by our pre-intervention theft proxy as most likely to engage in theft. Moreover, although registered buffet orders were higher in treated restaurants, the actual amount of food consumed—measured

by key breakfast ingredients in the restaurants' purchasing data—did not significantly differ from the control group. Based on these results and several robustness checks, we conclude that the only plausible explanation for the treatment effect is that treated servers could no longer bypass the POS system to pocket customer payments, while servers in the control group continued to do so.

The increase in registered breakfast buffets in response to our treatment translates to a substantial revenue gain for the firm (around 90 Euros per restaurant per morning shift), which is, by definition, exactly offset by a corresponding reduction in illicit income for servers who engage in theft. Our fine-grained data allow us to analyze how servers adjust to the loss of illicit income, looking at both turnover and income substitution.

Inspired by Coviello et al. (2022) and Krueger and Friebel (2022), which demonstrate that reducing an employee's regular monthly wage increases turnover, we study whether treated servers are more likely to quit the firm after the reduction of illicit income. We find that the treatment induced significantly higher attrition and turnover intention among male servers, but not among female servers.[1] In contrast, we find no treatment effects for attrition among bar/kitchen staff. This shows that increasing barriers to theft induces targeted sorting in our firm, where servers most likely to have engaged in theft are more likely to leave. As our exit interviews show that servers in our firm who quit generally stay in the restaurant sector, our treatment likely creates a negative externality by shifting the burden of misconduct onto competitors.

Apart from leaving the firm, servers can also substitute their lost illicit income through legal channels or alternative forms of employee theft. With respect to legal income, we find no treatment effects on generated tips or working hours, neither overall nor for male servers. For alternative forms of theft, we find some limited, mainly anecdotal evidence that male servers responded to the treatment by using alternative methods of theft. The only robust empirical finding, however, is that treated male servers who lost the most income (i.e., who were most likely to steal before the RCT) significantly increased their theft from *customers* instead of from the *firm*. Specifically, servers manipulate the transaction at the POS system by entering a fake item into the system right before payment—typically a low-priced, inconspicuous product such as a dip or ketchup that is unlikely to draw the customer's attention. They show this inflated bill to the customer on the POS screen, collect the higher amount, and immediately cancel the fake item afterward. This allows the server to pocket the price of the additional product while the system reflects only the legitimate order.

Thus, while the treatment protected the firm's direct revenue, it displaced misconduct toward customers, illustrating how a compliance measure can backfire by shifting theft toward a method with even greater reputational risk for the firm—since any discovery of these practices could erode customer trust and inflict long-term brand damage. This substitution into stealing from customers also affects the gender gap in illicit income: while our treatment eliminates the 10 percent gap originating from unregistered breakfast buffet orders, it simultaneously generates a 2 percentage

[1] Turnover intention was measured in a survey among servers during the RCT; these data were not collected for bar/kitchen staff.

point gap driven by theft from customers.

In sum, while formal hourly wages differ by only about 1 percent between male and female servers—and not at all within the same role—accounting for illicit income reveals a substantial gap in hourly earnings during the morning shift, which our treatment largely eliminates, leaving a small residual gap that stems from servers substituting into theft from customers.

A few months after the RCT concluded, the firm rolled out the treatment to all remaining restaurants. In the rollout period, registered buffet orders and server attrition no longer differed significantly between the former treatment and control restaurants.

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 document a substantial gender gap in 'illicit income' within a firm, to show that a compliance measure can eliminate this gap in one channel of misconduct, and to highlight how servers adapt by shifting to other forms of misconduct that open up a new gender gap.

Beyond the gender pay gap literature, our results also speak to the broader literature across criminology, sociology, and economics documenting that men are consistently overrepresented as perpetrators of crime. This pattern holds for violent offenses such as homicide and assault (Daly and Wilson 1988; Schwartz et al. 2009) and sex crimes (Bhuller et al. 2013), for property crimes including burglary and theft (Steffensmeier and Allan 1996; Campaniello and Gavrilova 2018), for corruption, bribe-taking, and tax evasion (Dollar et al. 2001; Swamy et al. 2001; Kastlunger et al. 2010), and for norm violations more broadly (Kruttschnitt 2013). Industry reports provide further evidence that employee theft and occupational fraud disproportionately involve male employees (Jack L. Hayes International 2023; Association of Certified Fraud Examiners 2024). We contribute to this literature in three ways: we provide new evidence on a previously underexplored margin—gender differences in employee theft; we show that raising the barriers to theft eliminates the resulting gender gap in unregistered buffet theft; and we document the resulting substitution into alternative forms of misconduct.

Our study also connects directly to the public economics literature on tax enforcement, which demonstrates that splitting transaction tracking across multiple parties is an effective tool to prevent tax evasion (see Slemrod 2019 for a review of this literature). While evasion is prevalent when a single person handles a transaction, introducing reporting by a third party (Kleven et al. 2011), cross-checking between different parties (Naritomi 2019), or verifiable paper trails between firms (Pomeranz 2015) systematically improves compliance. We provide causal RCT evidence that

extending this multi-party approach to the workplace reduces misconduct, and document the full behavioral response of affected servers to this change.

Extending this focus on compliance, our paper also contributes to the broader literature on workplace misconduct, which highlights the deterrent effect of higher relative wages (Chen and Sandino 2012), peer effects (Chan et al. 2021), customer characteristics (Balafoutas et al. 2013), and monitoring (Nagin et al. 2002). Most closely related to our work within this strand of literature is Pierce et al. (2015), who use a staggered rollout of an IT monitoring tool in US restaurants to show that improved monitoring reduces employee theft. The magnitude of our RCT effects is comparable to those documented by Pierce et al. (2015), suggesting that employee theft is a pervasive challenge in the restaurant sector more broadly. Our main contribution to this literature is to experimentally test a change to the underlying work process—rather than increased monitoring—that raises the barrier to misconduct, and to show that this has a direct, causal impact on gender gaps in illicit income: eliminating the existing gap tied to buffet theft while opening a new, smaller one as servers substitute into theft from customers. To our knowledge, no prior study has established this link between a compliance intervention and gender gaps.

2. Background

2.1. Study firm

Our study firm is a large all-day food and beverage chain that operates pub-style, full-service, sit-down restaurants, similar in format and price point to a mid-range casual-dining chain. It offers 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 70 restaurants across one European country that offer its all-you-can-eat breakfast buffet, most of them located in urban centers.[2] 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 equity share in the respective store. On average, an establishment employs 24 bar and kitchen staff members and 20 servers.

In the remainder of this paper, we focus on servers, as they are the group targeted by our treatment. Around two-thirds of servers are female, and they work full-time, part-time, or as marginal ("mini-job") employees, with full- and part-time contracts typically permanent. Wages are fixed at the national minimum wage—12.41 Euros per hour at the time of the study, the same rate that applies to workers across the economy. The only exception is a fixed monthly supplement of 300 Euros for shift supervisors, equivalent to about 1.73 Euros per hour for a full-time position. Pay is independent of how much servers collect in tips, and the firm does not pay any performance-based bonuses. Because pay is determined entirely by role, the gender gap in formal hourly wages is mechanically small: in the pre-RCT period, male servers earned on average 12.74 Euros per hour

[2]The firm operates ten additional restaurants that do not offer the breakfast buffet, due to limited physical space, and are therefore not part of our study.

and female servers 12.60 Euros, an unadjusted gap of 1.05 percent; conditional on restaurant and full-time status, the difference is 0.09 Euros (Table 1, Column 1). This gap arises solely because male servers are more often shift supervisors (18.8 versus 11.1 percent); conditional on role, the gap is zero.[3] Tips are customary but modest, typically amounting to rounding up the bill rather than a fixed percentage of it, and average around 3 Euros per hour.[4] We also find no significant gender difference in tips among servers in the pre-RCT period (Column 2 of Table 1).

2.2. Order process

In restaurants generally, the order process is central to operations, as essentially all revenue passes through it. Because theft is common within the hospitality industry (Pierce et al. 2015), this process typically involves more than one employee—from taking the order to preparing and delivering it—making it difficult for a single employee to bypass the system and keep a customer's payment without anyone noticing. This process is managed in restaurants through a point-of-sale (POS) system, which records each order and links it to the employee who placed it.

In our study firm, each transaction runs through a multi-party process. Upon taking an order, the server enters it into a handheld device, which transmits it to the bar or kitchen; bar and kitchen staff prepare what is registered in the POS, and the server picks up the prepared items and delivers them to the table. When the customer pays—roughly half pay in cash and half by card at the time of our intervention—the server settles the transaction in the POS and presents the receipt, typically displayed digitally on the handheld's screen, listing every product and the total amount due.[5] The multi-party process acts as an internal control: Since both the server and the bar and kitchen staff must handle the order, the opportunity for a single employee to bypass the system and deliver unregistered products is limited.

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, with restaurants selling an average of 61 breakfast buffets per day. The buffet consists of typical breakfast fare, including eggs, sausages, cheese, rolls, salads, vegetables and juice; however, hot beverages (such as coffee or tea) are excluded and must be ordered separately via the regular POS process. It is usually 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, eliminating the multi-party control and creating a straightforward opportunity for theft: a server can simply omit entering a buffet order into the

[3]The firm's personnel records contain each server's gender, contract type, and supervisor status, but not individual pay. We therefore impute hourly wages from the pay scheme—the minimum wage for all servers plus the supervisor supplement, converted to an hourly rate using the 173.33 monthly hours of a full-time position—for the 906 servers in our pre-RCT sample. Since pay is fixed by role, the only estimated quantity is the share of supervisors by gender. The unadjusted gap is statistically significant (standard errors clustered at the restaurant level, $p = 0.01$) but economically negligible.

[4]Individual servers keep the tips they receive but are required to share a small portion with bar and kitchen staff. This amounts to around two percent of their total generated revenue. The 3 Euros per hour take this already into account. 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). Moreover, we observe tips only in case of card payments. We extrapolate by assuming the same tip share for cash payments and card payment.

[5]Printed receipts are provided only upon explicit request.

POS system and keep the customer's payment for it. For cash payments, this leaves no trace at all; for card payments, the server books the excess amount as a tip.

2.3. Incentives for theft

Having the opportunity to steal, however, is not enough on its own: servers also need a strong incentive to exploit it. In our study firm, the expected costs of detection and punishment are low, while the potential financial gains are substantial.

Detecting theft in the first place is difficult, and even where the firm has gathered evidence, establishing legal proof is a separate, harder challenge. Management can only substantiate 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 employee theft is confirmed. These combined challenges mean that servers face very little risk of formal consequences.

Since servers earn minimum wage, the net gain from theft—which is tax-free and exempt from social insurance or pension contributions—can be substantial relative to their regular earnings. In line with Becker (1968) and the empirical findings of Jacob and Levitt (2003), this combination of weak sanctioning and considerable illicit gains keeps the expected costs of theft low relative to its expected gains, creating a strong incentive for misconduct.

2.4. Indicators of buffet payment theft

Interviews with managers and staff conducted prior to the RCT revealed a strong suspicion that servers frequently exploit the opportunity to bypass the POS system when handling breakfast buffet orders. Because such theft leaves little or no direct trace in the data, however, it cannot be observed directly. In this section, we therefore construct four indirect indicators of suspected buffet theft from the firm's granular transaction-level POS data for the pre-RCT period (from April 2024 until the start of the intervention). Our preferred and most comprehensive measure is the *buffet shortfall*, which we present first; the remaining three indicators—the *buffet–hot-beverage ratio*, *buffet cancellations*, and *implausibly large tips*—capture such buffet theft only under narrower conditions and serve mainly as complementary, corroborating evidence in later analyses.

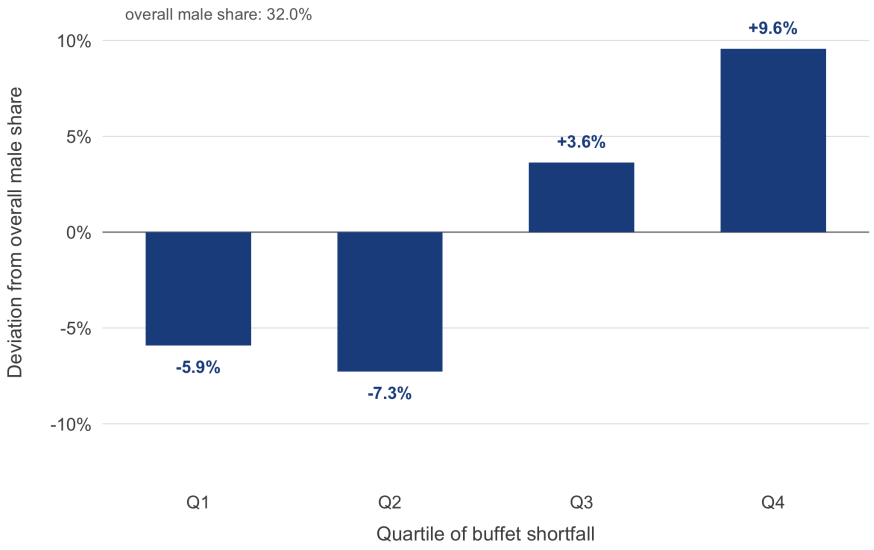
Buffet shortfall. We construct our preferred theft proxy by capturing how much each server deviates from expected buffet sales. Hot beverages—such as coffee and tea—are not included in the price of the breakfast buffet and must be ordered separately via the regular POS process, which maintains the multi-party control involving both servers and barkeepers. Because virtually all customers who order the buffet also order a hot beverage, this complementary good provides a reliable benchmark for actual buffet consumption at a given table. We therefore predict each server's buffet sales based on their sales of hot beverages, using a regression that includes a server

fixed effect to capture systematic deviations from this prediction.[6] To be precise, we estimate the following regression:

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

where s indexes restaurants, i servers, t days, w weekdays, and h hours. This regression estimates the number of breakfast buffets sold per hour conditional on the number of hot beverages sold, including time controls (restaurant × weekday × hour fixed effects), and most importantly a server fixed effect.[7] We recover 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 beverages sold and potential substitutes. Multiplying these fixed effects by –1 yields our buffet shortfall measure.

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 (32.0 percent), expressed in percent of the overall male share. Positive values indicate overrepresentation of male servers relative to the full sample, while negative values indicate underrepresentation.

Buffet-Hot beverage ratio. A simpler, closely related (purely descriptive) measure uses the same complementary-good logic without these additional controls: for each server, we compute the ratio of registered breakfast buffets to hot beverages ordered by customers at the same table. Systematically lower values for some servers indicate that the server is more likely not to have registered buffet orders in the POS system.

Buffet cancellations. A further method of theft involves initially entering the buffet order

[6]Similar approaches have been used in a labor productivity context by Mas and Moretti (2009) and in a crime context by Goncalves and Mello (2026).

[7]In some stores, a small share of revenue is also generated by other breakfast products. We therefore control for revenue from these products to improve the precision of our estimates.

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 system to allow servers to correct for mistakes in case they enter the wrong number or product when taking an order. However, an unusually high number of immediate cancellations by a given server is itself suspicious and therefore serves as an additional indicator of theft. This measure is narrower than the others, however, as it only catches servers who use this specific method rather than those who simply never register the order.

Implausibly large tips. A server seeking to keep the payment for an unregistered buffet order faces two scenarios, depending on how the customer pays. If the customer pays in cash, the server can simply pocket the money, leaving no trace in the POS data; if the customer pays by card, however, the payment must be recorded somewhere in the system, so the server instead can book it as a tip. Because tips under the prevailing social norm of rounding up the bill are typically modest, unusually large tips during the morning shift are therefore a sign of an unregistered buffet order. This measure is narrower than the others, however, since it is by construction only available for card payments and cannot identify theft in cash transactions.[8]

Appendix B.1 (Figure A.5) plots the pre-RCT distribution of each of these four indicators of buffet payment theft. In each case, the distributions reveal a distinct subset of servers with outcomes that are difficult to reconcile with ordinary service behavior alone—for instance, large negative buffet shortfalls, unusually low buffet-to-hot-beverage ratios, abnormally frequent immediate cancellations, or implausibly high tip shares—suggesting that these patterns reflect actual theft at the breakfast buffet rather than mere noise or heterogeneity in serving style.

Servers can also steal in ways unrelated to the breakfast buffet. Examples include collusive theft between barkeepers and servers, as well as implausibly large tips outside the morning shift. We return to these alternative methods in Section 5.4.3 (Appendix B presents the underlying pre-RCT evidence).

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

Before outlining our conceptual framework and our experimental design, we document a striking pattern in the pre-RCT period, which spans from April 1, 2024, to the start of the treatment: our main outcome of interest—the number of breakfast buffets each server registers in the POS system—varies systematically by gender. Female servers register on average 1.967 more buffets per day than male servers, a 9.3 percent gap (Table 1, Column 3)—a gap the firm's management has long interpreted as a productivity difference. Is this actually a productivity gap, however, or does it instead reflect differential rates of employee theft?

Our data provide strong support for this latter interpretation: female servers show a significantly lower buffet shortfall (that is, a lower propensity to fall short of predicted buffet sales) compared to

[8]The firm itself considers tips that systematically exceed 20 percent of the total bill 'implausible,' treating them as an indication that the server failed to register ordered products in the POS in order to generate additional income for themselves.

male servers (Column 4). This difference is driven by a gender skew at the upper tail of the buffet shortfall distribution: men are overrepresented in the top quartile by about 10 percent relative to their overall share in the sample (35.1 versus 32.0 percent, i.e., 3.1 percentage points; see Figure 1). Consistent with this result, we find that the average ratio of breakfast buffets to hot beverages is significantly lower for male servers (Table 1, Column 5). Our two additional theft indicators point in the same direction: female servers have fewer instant cancellations (Column 6), although this difference is not statistically significant, and a significantly lower sum of implausibly large tips (Column 7).

Taken together, these results provide a consistent picture: although the gender gap in formal hourly wages is only about 1 percent and entirely due to the share of shift supervisors (Column 1; Section 2.1), female servers register substantially more breakfast buffets than male servers, and all four theft indicators consistently point to higher rates of theft among male servers. Figure A.6 in Appendix B.1 shows that this difference is concentrated in the suspicious tail of each indicator: male and female servers behave similarly over most of the distribution, but the top quartile is markedly more pronounced among men. This suggests a substantial, theft-driven gender gap in illicit income. Our RCT provides causal evidence for this interpretation.

Turning to the other, buffet-unrelated forms of theft introduced above (Section 2.4), we find no evidence that women rely on these alternative methods more than men; if anything, men are also more likely to engage in them (see Appendix B for details). This suggests that the gender gap in illicit income we estimate in our RCT, based on breakfast buffet theft, if anything, represents a lower bound on the true gender gap.

2.6. Financial and legal consequences of theft

Employee theft is not only a source of gender inequality in illicit income—it is also costly for the firm itself, carrying severe financial and legal consequences. Unrecorded revenue has a direct, one-to-one impact on net income. Apart from eroding revenues, the systematic under-reporting 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: from the tax authorities' perspective, it has simply not paid taxes on that unreported income, and remains legally liable for the shortfall regardless of who pocketed the money. Thus, if employee theft is confirmed, the firm faces the risk of being forced to pay back taxes on the stolen amount along with fines for potential 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 firsthand: tax authorities issued inquiries after finding discrepancies between sales and purchasing records, placing the firm under suspicion of tax evasion.[9] Motivated by these risks, the firm implemented the intervention we study in this paper.

[9]These substantial costs also make it unlikely that management tacitly tolerates employee theft as an informal way of rewarding high-performing servers.

Table 1: Gender differences in earnings, productivity and theft proxies

	Earnings		*Productivity*	*Theft-proxies*			
	Hourly wage	Tips per day	No. Break-fast buffets	Buffet shortfall	Buffet-Coffee-Ratio	Cancel-ations buffet	Sum tips $> 20\%$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.088*	-0.917	1.967**	-0.300**	0.025**	-0.051	-0.701**
	(0.045)	(0.921)	(0.736)	(0.124)	(0.012)	(0.202)	(0.290)
Mean males	12.737	25.324	21.028	0.190	0.602	4.547	1.856
Observations/Workers	899	868	899	899	899	899	899
No. of Clusters	38	38	38	38	38	38	38
Controls	No	No	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 the server. All variables are aggregated to the server 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. Column 1 uses the formal hourly wage, imputed from the firm's pay scheme and personnel records: the statutory minimum wage of 12.41 Euros plus the fixed shift-supervisor supplement of 300 Euros per month (173.33 hours) for servers whose personnel record carries the shift-supervisor role during the pre-RCT period. Column 2 uses tips per day, extrapolated to cash payments as described in Section 2.1. Columns (6) and (7) include additional controls. In particular, Column 6 controls for Other Breakfast Revenue and captures revenue from breakfast items other than the buffet (i.e., substitute products). Column 7 controls for the average card share of payments of all transactions. Instant cancellations captures immediately cancelled breakfast buffets. The Buffet/Hot Beverage Ratio is defined as the ratio of breakfast buffets to hot beverages (in quantities). The Buffet shortfall corresponds to the server-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 servers in the estimation sample. The number of observations and the number of clusters are also reported. $^{*}< 0.1$, $^{**}< 0.05$, $^{***}< 0.01$

3. Conceptual framework

Our analysis is guided by a simple framework, which we state formally in Appendix A and summarize here. A server earns a fixed wage and can add to it in two ways: legally, by increasing effort to generate higher tips or by increasing working hours, or illegally, by keeping a customer's payment out of the firm's records. The framework rests on four assumptions. First, utility of income is increasing and strictly concave, as under standard risk aversion. Second, on each of the two margins, actions can be ordered from least to most costly, with the least costly action costless; a server therefore acts up to a cutoff at which the marginal cost of the next action equals the marginal utility of the earnings it generates—a cutoff that falls as total income rises. Third, following Becker (1968), the cost of an illegal action combines its moral cost with its expected sanction. In our setting, the cheapest form of theft is simply not registering a buffet order in the POS system, which requires no extra steps and leaves virtually no direct trace (Section 2.4); all other methods of theft are more cumbersome, riskier, and more severe (discussed in Section 5.4.3). Fourth, servers differ in integrity: for some, the moral cost of stealing is prohibitively high, so they never steal; for the others, it is finite and increasing, so they steal, starting with the cheapest method. Gender enters the framework in only one way: we assume that the share of servers who steal is larger among men,

in line with the evidence that men are more likely to engage in crime compared to women (Section 1).

This framework can provide a rationale for the pre-RCT patterns documented in Section 2.5 (Lemma 1). Servers who steal must earn more in total than servers who do not: otherwise, their marginal utility of income would be at least as high, so they would take at least as many legal actions—and earn illicit income on top, a contradiction. Their higher total income in turn lowers their legal cutoff, so their registered sales understate their true output twice over: they conceal part of the sales they generate, and they take fewer of the legal actions that generate sales and tips. Because men are more likely to steal, they should register fewer buffets and score worse on every theft indicator than women, even though they earn more in total—while formal, payroll-recorded earnings show no gender gap. One potential explanation for the productivity gap that management reads off its own records is thus a theft gap.

Our treatment (Section 4.1) removes the opportunity to pocket customer payments from the breakfast buffet (i.e., the easiest status quo opportunity), leaving servers who never steal unaffected. Registered buffets should therefore rise, while the number of buffets actually consumed should not—most strongly among servers whose pre-RCT buffet shortfall indicates that they relied on illegal income. For these servers, the lost illicit income cannot be fully recouped: if it were, their marginal utility of income would be back at its original level, both cutoffs back where they started, and no additional action taken. With lower income, marginal utility rises and both cutoffs move outward (Proposition 1). Affected servers should therefore respond on three margins: they may quit, as the value of employment has fallen for servers who stole—but only for them (Proposition 2); they may take more legal actions, generating higher tips or increasing working hours; and they may substitute toward illegal actions that were previously not worthwhile—by the cost ordering above, methods of theft that are more cumbersome, riskier, and more severe. Opportunities to increase legal income are limited in our setting: (i) the firm does not pay performance-based bonuses, (ii) it is unclear whether greater effort translates into higher customer tips, and (iii) working hours are fixed by the employment contract. In our setting, we therefore expect adjustment to occur mainly through exit or substitution toward alternative methods of theft.

Finally, because servers who never steal are unaffected, every treatment effect scales with the share of servers who steal and should therefore be larger for men than for women. The male–female difference in the treatment effect on registered buffets thus measures the pre-RCT gender gap in illicit income. The treatment closes this gap while the gap in payroll earnings remains at zero; if the escalation to more severe forms of theft concentrates among men, however, the gender gap in illicit income may re-open along a new margin.

4. The RCT

4.1. Design of our treatment

Our intervention is designed to mitigate employee theft stemming from unregistered breakfast buffet orders by making it substantially harder for servers to hand out buffet orders without registering

them in the POS. It leverages the fact that orange juice is included in the buffet: while juice remains self-serve in the control group, in the treatment group a customer's breakfast order triggers a different process—the server enters it into the POS system, which informs the bar keeper to prepare the juice; the server then picks it up from the bar and delivers it to the customer. In the control group, by contrast, the entire ordering process remains unchanged.

4.2. RCT implementation

The juice delivered from the bar in the treatment restaurants comes in a distinctive glass, purchased specifically for the intervention, that differs from the restaurant's standard glassware in several respects.[10] To implement this new ordering process in the RCT, we faced one practical challenge: producing and delivering the roughly 200 custom-made glasses required per treated restaurant. This challenge 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 restaurants offering the breakfast buffet to ask—without mentioning the RCT—if they were interested in "soon" introducing a new ordering process. Managers from 38 restaurants expressed interest and were enrolled in the RCT. Within this sample, the treatment and control groups were randomly assigned at the restaurant level.[11]

Second, the supplier's limited production capacity also meant that the custom-made glasses could not be delivered to all treatment restaurants simultaneously. The glasses were therefore delivered to the 20 randomly selected treatment establishments in two waves: half of the restaurants received the glasses on October 7th, 2024, the other half within the three weeks before. The delivery timeline was quasi-random, as the distribution was based purely on logistical factors.[12] Restaurants implemented the treatment immediately upon receiving the glasses.

Our main analyses focus on the RCT period, running from each restaurant's individual treatment start date through November 3, 2024 (results are similar when restricting the analysis period to October 7 onward—i.e., once all treatment restaurants had received their glasses; see Appendix B). Management communicated the new ordering process to its staff using a neutral framing, simply

[10] The glass incorporates the firm's logo and has a capacity of 0.3 liters, compared to the standard 0.2-liter glass used for self-served juice in the control group. Although a single glass is not refilled directly, customers in the treatment group can request additional juice at any time—each refill is simply prepared by the bartender and delivered by the server like a new order, consistent with the buffet's all-you-can-eat nature. According to the top management, customers on average drink one to two 0.2 liter glasses per buffet, for that reason, we decided to implement 0.3 liter glasses for our treatment. In our server survey and in Google reviews, we find no evidence that the perceived quality of the breakfast buffet offer suffered due to the treatment.

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

[12] The supplier delivered glasses to our study firm and other clients during the same logistics trips. Because the supplier lacked internal data from our firm, the delivery order was independent of restaurant characteristics.

describing it as a change to the buffet process without further detail.[13]

4.3. Randomization and descriptive statistics

Table 2 gives store-level descriptive statistics for all restaurants (Column 1) and separately for control and treatment restaurants (Columns 2 and 3) in the pre-RCT period (starting on April 1, 2024), along with standardized mean differences (see Imbens and Rubin 2015) and p-values from a regression of a treatment-group dummy on each outcome variable (Column 6). The 'all shifts' variables summarize data for the entire day, while the 'morning shifts' variables cover buffet hours (9:00 am–12:00 pm Monday–Saturday; 9:00 am–1:00 pm Sunday). We provide statistics for registered breakfast buffet orders (our key outcome variables), revenues (overall, and separately for beverages and other food), number of hot coffees and hot beverages, and the average number of servers, bar, and kitchen workers by shift. Similarly, the last panel of the table reports descriptive statistics, based on purchasing data, for three ingredients used exclusively for the breakfast buffet—bread rolls, sausages, and salmon—measured as the number of packages ordered and delivered to each restaurant. As all standardized mean differences but one remain below the threshold of 0.25 suggested by Imbens and Wooldridge (2009)—the exception is salmon, for which both group means are close to zero—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.

4.4. Server surveys

To measure how the intervention affected servers' behavior, perceptions, and turnover intentions, we surveyed all current servers at the beginning of November (response rate 28 percent). To understand the professional trajectories of former servers, including their subsequent employers and current jobs, we complement this with exit interviews conducted in late 2025 among 30 randomly selected servers who left the firm in the 12 months following the RCT (response rate 93 percent).

4.5. Ethics approval, statistical power, preregistration

We obtained approval from the Ethics 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, determining 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 transaction-level POS data on every customer order, the study is well-powered to identify statistically significant

[13] According to management, staff viewed the intervention as a measure to improve hygiene, reduce overcrowding at the buffet table, and lower costs by eliminating the unlimited offering. Consistent with the argument that servers did not perceive this as a theft-prevention measure, we found no treatment effect on perceived trust in management in our server survey (see Appendix D).

shifts in sales behavior. We registered the intervention in the AEA RCT Registry.[14]

5. Results of the RCT

5.1. Empirical strategy

In our main analysis, one observation in the data set corresponds to one server on a particular day in one morning shift in one restaurant. To analyze the treatment effect of increased barriers to employee theft, we estimate the following regression model:

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

where y_{it} represents the respective outcome variable; which is the number of sold breakfast buffets per shift in our baseline estimation. Servers are denoted by i on day t, while D_{it} is a treatment indicator, α_i captures restaurant/server fixed-effects, γ_t denotes time fixed-effects, and ϵ_{it} is the error term, which we cluster at the unit of randomization (Abadie et al. (2023)), the restaurant level. As the treatment was rolled out in a staggered fashion, 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.

5.2. Effect of the treatment on breakfast buffet sales

Our main outcome of interest is the number of breakfast buffets servers register in the POS system. As a first, descriptive step, we document how this outcome differs between treatment and control restaurants: treated restaurants register 70.79 breakfast buffets per day during the RCT on average, compared to 62.98 in control restaurants—a difference of 7.81, or 12 percent. Figure 2 shows that this difference reflects a broad shift across the entire distribution: a two-sample Kolmogorov-Smirnov test rejects the null hypothesis that the treatment and control distributions are equal (p-value < 0.001).

We next estimate the average treatment effect using Equation 2. Column 1 of Table 3 provides the result. We find an economically and statistically significant treatment effect: the mean number of registered breakfast buffet orders per day per server is 1.93 higher in treatment than in control restaurants, corresponding to an 8.7 percent difference. This effect is equivalent to 0.118 of the standard deviation of the number of registered breakfast buffets in the control group during the RCT period. These results are robust across a range of alternative specifications (see Appendix

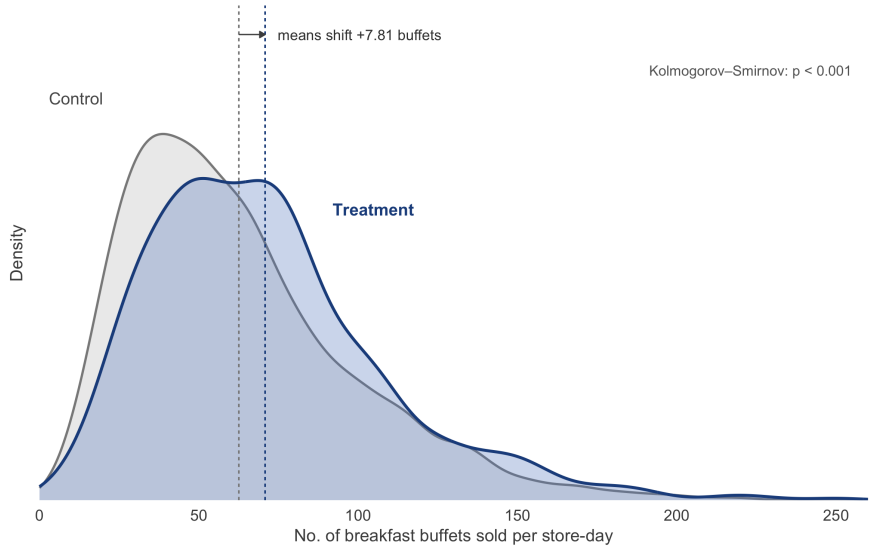
[14] Our pre-registration did not specify a focus on gender. The firm's personnel records were not, and could not easily be, directly linked to the POS data, so we did not initially expect to be able to measure differences in theft between female and male servers. Only after the pre-registration did we find a workaround to link the two data sources; once this was possible, examining gender was a natural choice given the well-documented literature showing that men are more likely to engage in criminal behavior than women (see, e.g., Steffensmeier and Allan 1996; Campaniello and Gavrilova 2018). Moreover, our gender gap appears consistently across four indicators for employee theft, as well as in the RCT treatment effect itself, making it unlikely that this pattern reflects specification searching along a single unregistered dimension. Correcting standard errors for multiple hypothesis testing using the Romano-Wolf procedure (Romano and Wolf 2005) yields unchanged conclusions.

Table 2: Balance table

	All	Control	Treatment	Absolute Δ	SMD Δ	p-value $\beta_T^{FE} = 0$
All shifts						
Sales	7024.69	6996.85	7050.03	53.18	0.017	0.935
	(3107.54)	(3163.32)	(3056.09)	(664.21)		
No. parties served	301.73	308.26	295.79	-12.47	-0.101	0.682
	(122.92)	(129.67)	(116.13)	(30.08)		
Morning shifts						
Sales and staffing (daily)						
Sales	1785.75	1758.57	1810.50	51.93	0.057	0.757
	(915.44)	(965.01)	(867.24)	(193.48)		
No. parties served	68.59	69.46	67.80	-1.66	-0.058	0.838
	(28.60)	(30.29)	(26.96)	(7.22)		
Quantity buffets	63.05	61.04	64.88	3.84	0.104	0.600
	(36.77)	(38.53)	(35.01)	(7.88)		
Sales buffets	860.19	829.58	888.07	58.49	0.106	0.555
	(549.98)	(565.78)	(533.75)	(108.65)		
Sales drinks	650.53	645.48	655.12	9.64	0.030	0.868
	(320.04)	(338.43)	(302.30)	(70.85)		
Sales other food	172.74	173.54	172.01	-1.53	-0.013	0.945
	(114.26)	(118.72)	(110.06)	(19.35)		
No. hot drinks	107.31	109.59	105.25	-4.34	-0.090	0.734
	(47.93)	(51.16)	(44.71)	(11.04)		
No. servers by shift	3.59	3.69	3.50	-0.19	-0.118	0.604
	(1.62)	(1.74)	(1.49)	(0.35)		
No. bar/kitchen workers by shift	4.75	4.75	4.76	0.01	0.005	0.959
	(1.89)	(1.90)	(1.89)	(0.40)		
Buffet ingredients (weekly packages ordered)						
Bread rolls	4.95	4.85	5.04	0.20	0.071	0.749
	(2.77)	(2.58)	(2.92)	(0.64)		
Sausages	0.85	0.75	0.93	0.18	0.223	0.189
	(0.79)	(0.74)	(0.83)	(0.13)		
Salmon	0.07	0.11	0.03	-0.08	-0.306	0.228
	(0.26)	(0.33)	(0.18)	(0.07)		
Observations (store-days)	6,826	3,253	3,573			
Observations (ingredient orders)	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, for both the treatment and the control group. The 'all shifts' variables summarize data for the entire day, while the 'morning shifts' 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.

Figure 2: Distribution of breakfast buffet sales by restaurant per day (kernel densities)



Notes: This graph shows kernel density estimates of the number of sold breakfast buffets by store per day. The gray area shows observations from the control group, the blue area those from the treatment group. Dashed vertical lines mark the group means; the arrow reports the difference in means between the two groups. A two-sample Kolmogorov–Smirnov test checking for equality of distributions yields a two-sided p-value smaller than 0.001.

B).[15]

[15]Specifically, three alternative estimators yield qualitatively similar results: a standard TWFE estimator restricted to the period after the staggered rollout was complete, a Poisson regression (Correia et al. 2020), and the difference-in-differences estimator of Callaway and Sant'Anna (2021). The latter is less precise, consistent with the weak serial correlation in our outcome, which favors regression-imputation estimators that draw on multiple pre-treatment periods over subgroup estimators that rely only on the last pre-treatment period (Harmon 2022). A placebo check further shows no significant treatment effects on revenue from food and beverages other than the breakfast buffet.

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 $H_0 : \tau = 0$	0.001	0.008	0.000
Romano-Wolf p-val	0.053	0.078	0.036
Observations	19,032	19,032	19,032
No. of Clusters	38	38	38

Notes: This table presents treatment effects on the number of registered breakfast buffets at the server-day level. Column 1 reports the pooled effect across all servers; Columns 2 and 3 report the effects estimated separately for female and male servers. The estimates are obtained using the imputation estimator of Borusyak et al. (2024), including server and time fixed effects; the gender-specific effects in Columns 2 and 3 are obtained from a single estimation that allows the treatment effect to differ by gender. *Control mean* is the average number of registered breakfast buffets per server-day in untreated server-days of the respective group. *p-val* H_0: $\tau = 0$ reports, for each column, the p-value of the test that the treatment effect equals zero; *Romano-Wolf p-val* reports the corresponding p-value adjusted for multiple hypothesis testing using the procedure of Romano and Wolf (2005). The p-value of the test that the treatment effects in Columns 2 and 3 are equal is 0.012. Standard errors clustered at the store level are reported in parentheses. $^{*}< 0.1$, $^{**}< 0.05$, $^{***}< 0.01$

As discussed in Section 2.5, female servers registered 9.3 percent more breakfast buffets than male servers prior to the RCT—an initial indication, together with our other theft proxies, that male servers were more likely to steal buffet payments. 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: female servers register on average 1.48 more breakfast buffets per day (corresponding to 5.9 percent, Column 2), while the treatment effect for male servers is around twice as large, at 2.89 more buffets per day (corresponding to 12.54 percent, Column 3). This difference in the treatment effects between genders is statistically 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. Valued at the average revenue per registered buffet of 13.65 Euros (see the notes to Figure 4), the difference in treatment effects of $2.89 - 1.48 = 1.41$ buffets per day implies that male servers pocketed about 19 Euros more per morning shift than female servers, or about 2.4 Euros per hour worked—roughly 19 percent of the hourly wage (assuming an eight-hour shift). Spread over the 16 hours of daily store operation (assuming no buffet theft outside the morning shift), this corresponds to a gender gap in hourly earnings of about 10 percent (see Appendix B.6 for the detailed calculation).

Taken together, these results show that the treatment substantially increased the number of registered breakfast buffets, with effects around twice as large for male servers as for female servers, causing a gender gap in illicit income. As laid out in our conceptual framework (Section 3), we interpret this increase in buffet orders in the treated restaurants as reflecting a reduction in employee theft. The higher number of registered buffets in treatment restaurants does not reflect additional

consumption relative to control restaurants, but rather sales that occur in control restaurants as well—there, however, the payment is pocketed by the server rather than recorded in the POS system.

5.3. Is the treatment effect really driven by theft?

This section presents a series of empirical facts that support the interpretation that the treatment reduced employee theft, and by doing so rules out alternative explanations such as customer demand.

5.3.1. Treatment effects on buffet inputs

Our first piece of evidence compares 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 appear in multiple menu items (e.g., tomatoes for both the buffet and burgers) or can be stored for several weeks (e.g., butter). We therefore focus on bread rolls, sausages, and salmon as proxies for actual consumption: they are used exclusively for the breakfast buffet, consumed by most guests, and cannot be stored long-term. As shown in Table 4, we detect no significant treatment effect on the number of rolls, sausages, or salmon ordered. This suggests that physical consumption is unaffected by the treatment, supporting the view that the treatment increases not the number of buffets ordered and consumed, but the number registered in the POS.

Table 4: Treatment effects on purchases of buffet ingredients

	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. $^{*}< 0.1$, $^{**}< 0.05$, $^{***}< 0.01$

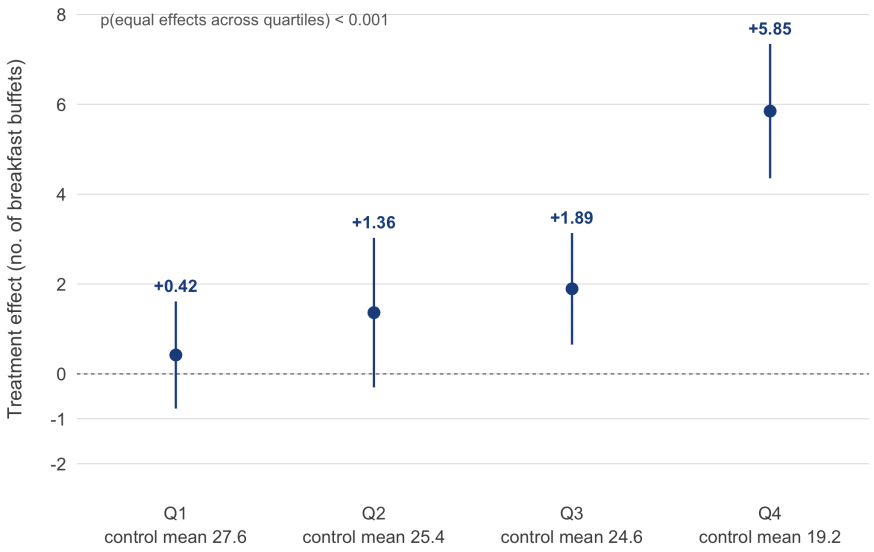
5.3.2. Treatment effects by pre-RCT buffet shortfall

Our second piece of evidence uses the buffet shortfall indicator introduced in Section 2.4, which captures whether a server systematically registered fewer breakfast buffets than predicted by their

sales of complementary and substitute products (e.g., hot beverages).

Figure 3 presents treatment effects by quartile of this pre-intervention buffet shortfall.[16] $Q1$ corresponds to the 25 percent of servers with the lowest value of λ_i—those least likely to have engaged in employee theft prior to the RCT—while $Q4$ contains the 25 percent with the highest value, and thus the highest likelihood of having done so; $Q2$ and $Q3$ are constructed analogously. As the figure shows, treatment effects are largest for servers in $Q4$, that is, for those most likely to have failed to register breakfast buffet orders before the intervention.[17]

Figure 3: Treatment effect by buffet shortfall



Notes: This figure shows point estimates and 95 percent confidence intervals of the impact of the treatment on the number of sold breakfast buffets by server by quartile of buffet shortfall. For instance, $Q4$ shows the aggregated treatment effects of the 25 percent of servers proxied as having the highest likelihood of engaging in illegal behavior due to high buffet shortfall prior to the RCT. Below each quartile, the average number of registered breakfast buffets in untreated server-days of that quartile is reported. The annotation reports the p-value of the test that the treatment effects are equal across all four quartiles. Standard errors are clustered at the store level.

Because men are overrepresented in $Q4$, the quartile of servers with the highest pre-intervention theft propensity (Figure 1), this raises the question of whether the gender gap in the main treatment effects on registered breakfast buffets documented above simply reflects this compositional difference. To study this, we compare treatment effects by gender within each quartile of the buffet shortfall measure. Table 5 shows that, within $Q4$, we find significant treatment effects for both male and female servers, with a somewhat larger effect for men—indicating that, even among servers with a similarly high propensity to steal, men steal more. Taken together with the compositional imbalance documented above, this suggests that differences in the propensity to steal at the outset—rather than differential behavior within a given propensity group—are the main driver of the overall gender gap in the treatment effect.

[16] See Table A.12 in Appendix B for the corresponding regression table.

[17] 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.

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 server, by quartile of the buffet shortfall measure and by gender. Columns 1 and 2 ($Q1$–$Q3$) show the treatment effects for female and male servers among the 75 percent of servers proxied as having a relatively lower likelihood of engaging in illegal behavior prior to the RCT; Columns 3 and 4 ($Q4$) show the corresponding effects for the 25 percent of servers 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 shows, for each quartile group, the p-value for the test of equality of the treatment effects between female and male servers. Standard errors clustered at the store level are reported in parentheses. $^{*}< 0.1$, $^{**}< 0.05$, $^{***}< 0.01$

5.3.3. Treatment effects on buffet payment theft indicators

As a final piece of evidence, we examine treatment effects on the three remaining indicators of buffet payment theft introduced in Section 2.4—buffet cancellations, the buffet–hot-beverage ratio, and implausibly large tips—again separately for the top quartile of the pre-RCT buffet shortfall distribution and the remaining sample (Table 6). Among servers in $Q4$, the treatment significantly reduces the number of buffet cancellations by about 23 percent (Panel A) and raises the buffet–hot-beverage ratio by about 17.5 percent (Panel B), whereas the effects for $Q1$–$Q3$ are small and insignificant. The sum of implausibly large tips, by contrast, does not decline (Panel C). This indicator is arguably the noisiest of the four: it is observed only for card payments and also captures other methods of theft, to which servers may substitute once buffet theft becomes harder (Section 5.4.3).

5.4. How servers respond to the loss in illegal income

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 B.6). This revenue gain is, by definition, exactly offset by a corresponding reduction in illicit income for servers who engage in theft.

Figure 4 translates these treatment effects into net losses of illicit income per hour worked, by quartile of the pre-RCT buffet shortfall introduced in Section 2.4 (Panel a) and by gender (Panel b). For servers in $Q4$, the loss amounts to about 10 Euros per hour worked, or 80 percent of the gross hourly wage, compared with less than 1 Euro in $Q1$; for male servers it is about twice as large

Table 6: Treatment effects on complementary indicators of buffet theft

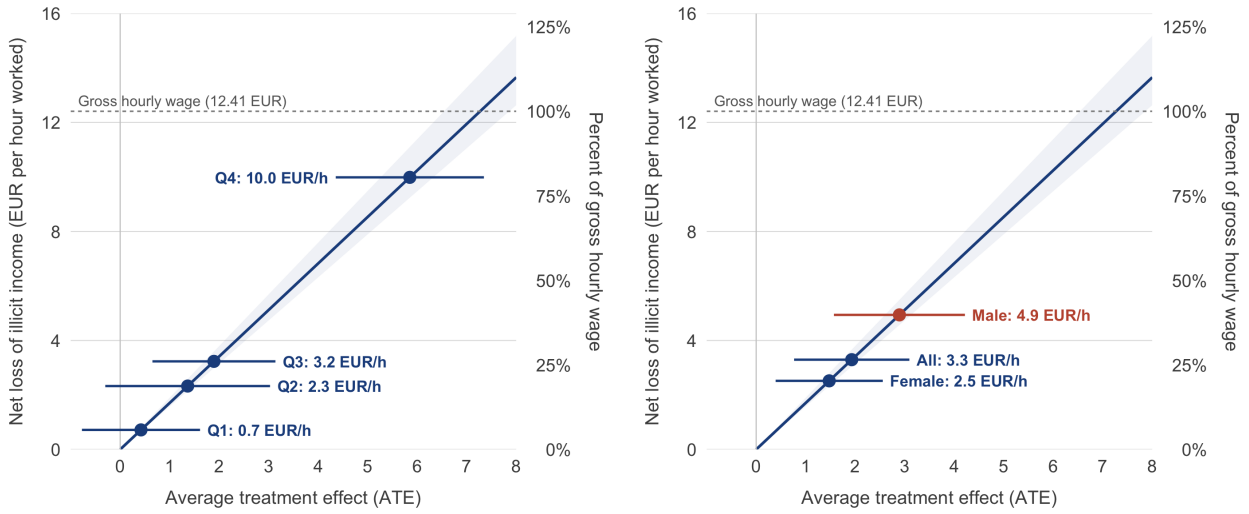
Panel A: Instant cancellations buffet		
	Q1–Q3	Q4
	(1)	(2)
Treatment	-0.114	-0.467***
	(0.093)	(0.174)
Control mean	1.732	2.001
p-val $H_0 : \tau_{Q4} = \tau_{Q1–Q3}$	0.050	
Observations	16,091	
No. of Clusters	32	
Panel B: Food-Hot Drink-Ratio		
	Q1–Q3	Q4
	(1)	(2)
Treatment	0.009	0.085***
	(0.015)	(0.023)
Control mean	0.643	0.486
p-val $H_0 : \tau_{Q4} = \tau_{Q1–Q3}$	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 $H_0 : \tau_{Q4} = \tau_{Q1–Q3}$	0.513	
Observations	19,032	
No. of Clusters	38	

Notes: This table shows the impact of the treatment on complementary indicators of buffet theft, by quartile of the buffet shortfall measure. Column 1 ($Q1$–$Q3$) shows the treatment effect for the 75 percent of servers proxied as having a relatively lower likelihood of engaging in illegal behavior prior to the RCT; Column 2 ($Q4$) shows the treatment effect for the 25 percent of servers 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 each panel shows the p-value for the test of equality of the treatment effects between $Q4$ and $Q1$–$Q3$. Standard errors clustered at the store level are reported in parentheses. $^{*}< 0.1$, $^{**}< 0.05$, $^{***}< 0.01$

as for female servers (4.9 versus 2.5 Euros per hour).

In the following subsections, we examine three ways in which servers may adjust their behavior in response to this loss in illicit income, as implied by our conceptual framework (Section 3): (i) the reduction in earnings capacity may increase turnover, as affected servers turn to better outside options (e.g., other restaurants in the sector). But even servers who do not quit may adjust their

Figure 4: Net loss of illicit income per hour worked, by average treatment effect (ATE) on the number of registered breakfast buffets per server and morning shift



(a) ATE by quartile of pre-RCT buffet shortfall

(b) ATE by gender

Notes: This figure shows the net loss of illicit income per hour worked that the treatment implies for a server, as a function of the average treatment effect on the number of registered breakfast buffets per server and morning shift. The solid line converts a given treatment effect into Euros per hour worked: each additional registered buffet corresponds to one buffet payment that is no longer pocketed, valued at the average revenue per registered buffet in the estimation sample (13.65 Euros; the shaded band spans the weekday and weekend averages of 12.64 and 15.18 Euros). Breakfast buffets are only served during the morning buffet window (four hours on weekdays and five hours on weekends, 4.4 hours on average across server-days), whereas a regular shift lasts eight hours and offers no opportunity for buffet theft outside this window; the loss per morning shift is therefore divided by eight hours. The dashed line marks the gross statutory minimum wage of 12.41 Euros per hour (Section 2.1); the right axis expresses the loss as a percentage of this wage. Since illicit income is untaxed whereas the wage is gross, these percentages understate the loss relative to net formal earnings. Dots place the estimated treatment effects on the line: Panel a by quartile of the pre-RCT buffet shortfall, with $Q1$ the 25 percent of servers least likely and $Q4$ the 25 percent most likely to have engaged in buffet theft before the RCT (estimates from Table A.12); Panel b for all servers and separately by gender (estimates from Table 3). Horizontal bars are 95 percent confidence intervals of the treatment effects, based on standard errors clustered at the restaurant level (N = 19,032 server-days, 38 clusters).

behavior by (ii) substituting into legal activities, for example by increasing effort to generate higher tips or working more overtime, or (iii) substituting into alternative forms of theft.

5.4.1. Do servers quit the firm more often?

We first examine whether the reduction in illicit income induced by the treatment affects server attrition. To do so, we construct a binary indicator for whether a server quit the firm in a given month, based on the firm's personnel files. We then estimate the treatment effect on this indicator

over the period from the start of the RCT through the end of December 2024.[18] Panel A of Table 7 presents these results overall (column 1) and separately by gender (columns 2 and 3). We find that the treatment increases attrition among male servers, but has no discernible effect on female servers. The magnitude is economically substantial: men in the treatment group are 58 percent more likely to quit than those in the control group.[19]

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 server 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. $^{*} < 0.1$, $^{**} < 0.05$, $^{***} < 0.01$

[18] Our treatment ran until November 3, 2024, after which we made small changes to it in the treatment restaurants—adding a hot beverage to the buffet order and raising its price slightly (see Section 6)—before starting to roll it out to the remaining restaurants on January 1, 2025. Our attrition analysis covers the period from the start of the RCT through December 31, 2024, because (i) the treatment remained virtually identical from the servers' perspective during this time (only the customer interface changed slightly; daily work and incentives were unaffected, making differential effects on turnover highly unlikely), (ii) servers commonly time their resignations around the end of the calendar year, and (iii) turnover lags behind quitting decisions: with a four-week notice period, servers who decided to quit in response to the treatment may not formally leave until weeks later. Our main qualitative findings are robust to restricting the analysis to the primary RCT period (i.e., through November 3, 2024) or expanding it through January 31, 2025, since the four-week notice period means these quits still largely reflect decisions made before the January 1 rollout.

[19] Ideally, we could also study whether servers most likely to have increased their income through breakfast buffet theft in the pre-RCT period—i.e., those in the fourth quartile of our buffet shortfall measure—are driving the results. However, we do not have sufficient statistical power to conduct such an analysis in an appropriate way.

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 little direct opportunities to steal via unregistered breakfast buffets. We find no significant treatment effect on the attrition of these groups, either overall or separately by gender.[20]

Finally, we examine whether actual quits are mirrored by subjective turnover intentions, using data from a server survey conducted at the end of the RCT, in which servers reported their likelihood of leaving the firm. Panel B of Table 7 reports the results.[21] Consistent with the results on actual turnover, we find no treatment effect for female servers; male servers in the treatment group, by contrast, report turnover intentions that are 0.46 standard deviations higher than those of male servers in the control group.[22]

Taken together, our results show that male servers in the treatment group exhibit both significantly higher turnover intentions 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 servers 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 servers at competing firms within the local labor market.[23]

5.4.2. Do servers generate more legal income?

Servers may respond to the reduction in illicit income by increasing their effort to generate more legal income—either during their existing working hours or by working more hours. Because the firm pays no performance-based bonuses (Section 2.1), the only way more effort within existing hours can translate into higher legal earnings is through tips—consistent with Pierce et al. (2015),

[20]Barkeepers do have more work in the treated restaurants during the RCT, but the effect is negligible, since filling juice glasses added only a few seconds per glass. In line with this negligible increase, when asked whether their bar and kitchen colleagues experienced more stress, servers in treatment and control restaurants reported no significant difference (see Appendix D); barkeepers' and kitchen staff's attrition rates also remained stable across groups. Taken together, this evidence suggests that the intervention did not meaningfully increase job disutility or alter the non-pecuniary aspects of barkeepers' and kitchen staff's daily work.

[21]To measure their intention to quit, servers were asked: (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.

[22]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. Beyond direct dismissals, it is also 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 servers.

[23]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'—tolerating such misconduct would be a high-risk strategy in practice, 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 also challenging in our setup, since standard performance metrics could themselves be correlated with theft. We therefore use the average plausible tip (controlling for gender) as our preferred productivity measure, as it reflects service quality from the customer's perspective. In our RCT, treatment effects on registered breakfast buffet orders are similar for servers with above- and below-median average tips in the pre-RCT period (see Appendix B), suggesting that more productive servers do not steal more.

who find that a theft-reducing IT intervention increased average tips in restaurants. Our server survey provides suggestive evidence of such an effort response: we find a significant treatment effect on self-reported activities aimed at generating higher tips (Appendix B.7). However, it is unclear whether these efforts actually translate into higher tips from customers (e.g., better service may go unnoticed, or overly persistent requests for tips may backfire). Indeed, we find no treatment effect on (plausible) tips, neither overall nor for male servers, nor for the subsample of servers most likely to have increased their income through breakfast buffet theft in the pre-RCT period, i.e., those in the fourth quartile of our buffet shortfall measure (Appendix B.7).

An alternative way for servers to substitute into legal earnings is to work longer hours. A literature on income targeting and reference points (Camerer et al. 1997, Fehr and Goette 2007, Crawford and Meng 2011, Thakral and Tô 2021, Kukavica et al. 2022, Duong et al. 2023) provides evidence that individuals adapt their labor supply to avoid falling below income reference points shaped by previous earnings. However, the scope for this is limited in our setup, as hours are fixed by the employment contract and overtime is only permitted when customer demand at the end of a shift is high. Consistent with this institutional setting, we find no treatment effect on working hours.

Taken together, these results suggest little to no systematic substitution into legal earnings.

5.4.3. Do servers substitute into alternative forms of theft?

Servers may also respond to the reduction in illicit income by substituting into alternative forms of theft that the treatment does not directly target. As briefly mentioned in Section 2.4, several such alternatives exist. Following discussions with the firm's management, previous research (Chen and Sandino 2012, Pierce et al. 2015, Chan et al. 2021, Brady et al. 2012), industry reports (National Restaurant Association 2025), and AI-based brainstorming, we identify the methods we consider most likely to occur in our setting. These methods vary in how difficult they are to implement, as well as in their profitability and probability of detection.

i) Kids breakfast buffet. In addition to the regular breakfast buffet, the firm offers a children's breakfast buffet priced at half the regular price. When serving an adult customer, a server can register the order in the POS as a children's buffet—handing out the accompanying orange juice glass to maintain appearances—while still charging the customer the full regular price. The server pockets the difference between the regular and the children's price directly if the customer pays in cash, or books it as a tip if the customer pays by card.

This method is relatively straightforward to implement and highly profitable. The probability of detection is nonetheless non-negligible: because the children's buffet is rarely booked, any spike in orders stands out immediately in the POS data.

ii) Direct theft from the bar or kitchen. A server can leave an item off the order in the POS while registering the rest of the order normally, surreptitiously take the item from the bar or kitchen, and still charge the customer its regular price. The server then pockets this payment directly if paid in cash, or books it as a tip if paid by card.

This method is less straightforward: the server must already decide, when taking the customer's order, which item to remove from the bar or kitchen, and must then retrieve it secretly and within a short time window, without being noticed. Taking an item inconspicuously from the bar or kitchen is, if anything, only feasible for bottled beverages and a few other products that are not freshly prepared; since such products also tend to be comparatively cheap, this further limits the method's profitability. The probability of detection is relatively high, since servers are generally not permitted behind the bar or in the kitchen.

iii) Collusion with bar and kitchen staff. A server and a bar or kitchen employee may collude to bypass the POS system entirely: the order is never entered, but the bar or kitchen employee prepares the item anyway, and the server delivers it to the customer, who pays the regular price. The server pockets this payment directly if the customer pays in cash, or books it as a tip if paid by card, sharing the proceeds with the colluding bar or kitchen employee.

This method is difficult to set up, as it requires stable coordination between a server and a bar or kitchen employee; once such cooperation is established, however, execution becomes easy, profitability can be high, and the probability of detection is comparatively low. The arrangement collapses, however, as soon as one of the two colluding employees leaves the firm.

iv) Theft from the customer. A server may enter a fake item—typically a low-cost product such as a dip or ketchup—into the POS immediately before payment, show the resulting, inflated bill to the customer on the POS screen, and collect the higher amount. The server then cancels the fake item immediately after payment, so that the POS reflects only the legitimate order, and pockets the price of the fake item directly if the customer paid in cash, or books it as a tip if the customer paid by card.[24]

This method has low implementation complexity. Profitability per transaction is rather low, but the method can be applied to many customers. It carries only moderate detection risk—particularly when the fake items are relatively cheap—since servers can always offer the customer a plausible excuse for why a product appears incorrectly on the POS screen, making it unlikely that the customer will report it. Detection by management was similarly limited: management was unaware of how to measure this method, and we developed a workaround in the firm's IT system to identify it after the RCT.

Overall, the four methods trade off differently along these dimensions: the kids breakfast buffet method is easy to execute and highly profitable but leaves a visible trace; direct theft from the bar or kitchen is difficult, only modestly profitable, and easily detected; collusion is hard to establish but, once running, is both profitable and hard to detect; and theft from the customer is easy to

[24] A related method uses the same cancel-after-payment mechanic as method (iv) but steals from the firm rather than the customer, and is limited to shift supervisors, who have supervisor-level access to the POS system: they enter a real product as usual, letting it pass through the standard multi-party process via the bar or kitchen, but then cancel the order after picking it up and still collect the customer's regular payment—pocketing it directly if paid in cash, or booking it as a tip if paid by card.

execute and hard to detect, but only modestly profitable per transaction.[25]

We use POS data to construct indicators for each of these methods. The number of registered children's buffets serves as a direct indicator for the *kids breakfast buffet* method. The *kids breakfast buffet*, *direct theft from the bar or kitchen*, and *collusion with bar and kitchen staff* methods—and, in part, also *theft from the customer*—all leave a common trace in the POS data: implausible tips, the same indicator that management uses as its general anti-theft measure. For *collusion with bar and kitchen staff* specifically, we go further and test for collusion directly by examining whether particular server–bar/kitchen employee pairs are systematically associated with higher implausible tips (see Appendix B.7 for details). Finally, we measure *theft from the customer* directly in the data by identifying products that are entered into the POS and canceled again shortly before the payment is booked (see Appendix B.7 for the exact definition).

Table 8 presents treatment effects on our indicator of *theft from the customer*, separately by gender and by pre-RCT theft propensity. For servers in the first three quartiles of the buffet shortfall distribution, we find no significant treatment effect for either gender (Columns 1 and 2). We then restrict the sample to the servers most likely to have increased their income through breakfast buffet theft in the pre-RCT period—those in the fourth quartile of our buffet shortfall measure. Recall that these are the servers for whom we estimate a pre-RCT income increase of about 80 percent through breakfast buffet theft alone. Within this group, the treatment significantly increases theft from customers for male servers, by about 25 percent relative to the control mean (Column 4), but not for female servers (Column 3); the difference between the two effects is statistically significant ($p = 0.026$). Appendix B.7 reports the corresponding results for the indicators of the other three methods.

To dig deeper into this, we analyze in Appendix B.7 the fake items servers use when committing *theft from the customer*. These items are typically inexpensive, mostly condiments and side dishes.

Thus, while the treatment protected the firm's direct revenue, it displaced misconduct toward customers—introducing a reputational risk, since any discovery of these practices could erode customer trust and inflict long-term brand damage. This has two important implications. First, while our treatment closes the 10 percent gender gap in illicit income originating from unregistered breakfast buffet orders, it simultaneously opens up a 2 percentage point gender gap driven by theft from customers. Second, these results illustrate how a compliance measure can have unintended consequences: by restricting the theft it targets, it may inadvertently shift misconduct toward methods that pose even greater reputational risks for the firm.

[25] In principle, method (ii) could also be combined with other methods. For example, a bar keeper could allow a server to surreptitiously take an item from the bar (method ii) and split the proceeds with the server, effectively combining methods (ii) and (iii). Alternatively, a server could leave an item off the order in the POS, then enter it immediately before payment and cancel it immediately afterward, combining methods (ii) and (iv). We consider these combinations unlikely, since method (ii) is complicated to execute, not very profitable, and carries a high probability of detection—all of which would still apply to the combined method. With respect to the combination of methods (ii) and (iv), this is corroborated by the results in Appendix B.7, which show that the fake items servers use under method (iv) are almost exclusively low-cost condiments—items whose physical removal from the bar or kitchen would hardly be worth the risk.

Table 8: Treatment effects on theft from the customer

	No. cancellations shortly before payment			
	Q1–Q3		Q4	
	Female (1)	Male (2)	Female (3)	Male (4)
Treatment	-0.045	-0.020	-0.033	0.094**
	(0.041)	(0.041)	(0.060)	(0.043)
Control mean	0.292	0.312	0.267	0.376
p-val $H_0 : \tau_{Male} = \tau_{Female}$	0.504		0.026	
Observations	38,317			
No. of Clusters	38			

Notes: This table presents treatment effects on the number of cancellations made by waiters up to 45s before payment. Estimates can be interpreted as marginal effects. Standard errors are reported in parentheses and clustered at the store level. $^{*} < 0.1$, $^{**} < 0.05$, $^{***} < 0.01$

6. Post-RCT period: Modified treatment and roll-out

Following the RCT period, the firm modified the intervention slightly in the treatment restaurants between November 4 and December 31, 2024: hot beverages, which were excluded from the buffet during the RCT, were bundled into the buffet price, accompanied by a one-euro price increase. This modification increased the treatment intensity, as the barkeeper's involvement shifted from preparing a single item (juice) to preparing two items (juice and hot beverages). The results for this modified treatment are qualitatively the same as our main RCT findings: registered buffet orders remain higher in treatment than in control restaurants, and the effect is again concentrated among male servers. Because the modification combined the inclusion of a new product with a price increase, however, we cannot isolate the effect of the modified service process from the change in pricing. Detailed results are reported in Appendix B.8.

At the beginning of January 2025, the firm began rolling out the treatment to all restaurants, including both the control restaurants of the RCT and those that had initially not been interested in adopting a new ordering process for the buffet. Because the supplier's production capacity was limited, the CEO let each restaurant choose whether to introduce the modified buffet, the original juice process, both, or neither, immediately or within a few weeks; adoption between January and April 2025 was therefore staggered and potentially endogenous, and we do not attempt to identify treatment effects for this period. From May 2025 onward, all restaurants used the same ordering process: juice and hot beverages are both served from the bar, and the buffet price is one euro higher than at baseline. In this period, the former treatment and control restaurants converge: registered buffet orders no longer differ significantly between the two groups, and the same holds for server attrition, where the differential quit rate among male servers observed during the RCT disappears. Detailed results are reported in Appendix B.8.

7. Conclusion

In this paper, we study gender gaps in illicit income within a firm in the food service industry, a sector that alone employs around 10 percent of the workforce in the US as well as the EU (U.S. Bureau of Labor Statistics 2025, Eurostat 2024). Using a large-scale RCT conducted within a major restaurant chain, we show that a significant portion of the gender pay gap can be hidden within such illicit income—informal earnings derived from workplace misconduct that remain invisible in formal payroll data. We find that a single, compliance-enhancing measure closes this gap, though a new one emerges as male servers adapt their behavior.

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. Our intervention, which involved multiple employees in the ordering process, thus offers a more accurate measure of servers' true productivity than the firm's own sales records had suggested. This implies that unaudited performance metrics can systematically confound misconduct with genuine ability differences, with consequences that may extend beyond pay to decisions such as promotions or internal recognition.

The elimination of this illicit income led to significant changes in the workforce and its behavior. We document a server 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 servers out of the firm. Because these departing servers 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. This implies that firm-level compliance measures, while effective at reducing misconduct within the adopting firm, have a weaker effect at the level of the sector as a whole—since displaced servers can simply move to competitors that have not yet adopted similar safeguards. This suggests a role for industry associations or regulators in helping standardize and accelerate the diffusion of such measures across the sector.

Among male servers who remained, those who had lost the most illicit income increasingly turned to stealing directly from customers rather than from the firm. Because this pattern was concentrated among men, it reopens a gender gap in illegal income—this time borne by customers rather than by the firm. This shift also exposes the firm to reputational risk, illustrating how a compliance measure can backfire by displacing misconduct toward a channel with potentially greater long-term costs. This implies that firms adopting theft-reducing measures should anticipate, measure, and address such substitution effects, rather than treating a reduction in the targeted misconduct as evidence that overall misconduct has declined. Our findings also imply that firms would have an economic incentive to staff payment-handling roles primarily with women.

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. This implies that researchers and policymakers assessing gender equity

based on payroll parity alone should be cautious in any sector where employees directly handle customer payments, as such data may mask meaningful disparities in total economic returns. 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.

Our study shows a gender gap in illicit income, but it does not resolve the broader, long-standing question in the literature of why men are, on average, more likely to engage in theft and other forms of criminal behavior than women (Section 1). Proposed explanations include gender differences in risk, competitive, or social preferences (Croson and Gneezy 2009, Dohmen et al. 2011, Buser et al. 2014, 2026), the role of peer effects in propagating and sustaining misconduct within work groups (Chan et al. 2021), differences in self-control (Gottfredson and Hirschi 1990), biological and hormonal factors (Sapienza et al. 2009), and differential responses to strain (Broidy and Agnew 1997), among others. Our large-scale RCT provides new causal evidence on the extent of this gap in a real labor-market setting, but isolating which of these mechanisms drives it remains a task for future research.

References

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.

Association of Certified Fraud Examiners (2024). Occupational fraud 2024: A report to the nations. `https://legacy.acfe.com/report-to-the-nations/2024/`. Accessed: 2026-08-30.

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.

Bhuller, M., T. Havnes, E. Leuven, and M. Mogstad (2013). Broadband internet: An information superhighway to sex crime? *The Review of Economic Studies 80*(4), 1237–1266.

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

Bloom, N., J. Liang, J. Roberts, and Z. J. Ying (2015). Does working from home work? Evidence from a Chinese experiment. *The Quarterly Journal of Economics 130*(1), 165–218.

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.

Broidy, L. and R. Agnew (1997). Gender and crime: A general strain theory perspective. *Journal of Research in Crime and Delinquency 34*(3), 275–306.

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

Buser, T., M. Niederle, and H. Oosterbeek (2026). Can competitiveness predict education and labor market outcomes? Evidence from incentivized choice and survey measures. *The Review of Economics and Statistics 108*(3), 755–773.

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

Camerer, C., L. Babcock, G. Loewenstein, and R. Thaler (1997). Labor supply of new york city cabdrivers: One day at a time. *The Quarterly Journal of Economics 112*(2), 407–441.

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.

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.

Crawford, V. P. and J. Meng (2011). New york city cab drivers' labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income. *American Economic Review 101*(5), 1912–1932.

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

Daly, M. and M. Wilson (1988). *Homicide*. Hawthorne, NY: Aldine de Gruyter.

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

Dollar, D., R. Fisman, and R. Gatti (2001). Are women really the "fairer" sex? Corruption and women in government. *Journal of Economic Behavior & Organization 46*(4), 423–429.

Duong, H. L., J. Chu, and D. Yao (2023). Taxi drivers' response to cancellations and no-shows: New evidence for reference-dependent preferences. *Management Science 69*(1), 179–199.

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

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.

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.

Gottfredson, M. R. and T. Hirschi (1990). *A General Theory of Crime*. Stanford, CA: Stanford University Press.

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

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

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

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.

Kastlunger, B., S. G. Dressler, E. Kirchler, L. Mittone, and M. Voracek (2010). Sex differences in tax compliance: Differentiating between demographic sex, gender-role orientation, and prenatal masculinization (2D:4D). *Journal of Economic Psychology 31*(4), 542–552.

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.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica 79*(3), 651–692.

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

Kruttschnitt, C. (2013). Gender and crime. *Annual Review of Sociology 39*, 291–308.

Kukavica, A., S. McKenna, M. Shum, M. K. Chen, and C. Camerer (2022). Beyond taxis: Reference-dependence in rideshare drivers' labor supply. *Available at SSRN 4205411*.

Mas, A. and E. Moretti (2009). Peers at work. *American Economic Review 99*(1), 112–145.

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.

Naritomi, J. (2019). Consumers as tax auditors. *American Economic Review 109*(9), 3031–3072.

National Restaurant Association (2025). 2025 state of the restaurant industry. `https://restaurant.org/research-and-media/research/research-reports/state-of-the-industry/`. Accessed: 2026-08-18.

NRF (2025, October). The impact of retail theft & violence 2025. `https://nrf.com/research/the-impact-of-retail-theft-violence-2025`. Accessed: 2026-03-01.

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.

Pomeranz, D. (2015). No taxation without information: Deterrence and self-enforcement in the value added tax. *American Economic Review 105*(8), 2539–2569.

Romano, J. P. and M. Wolf (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association 100*(469), 94–108.

Sapienza, P., L. Zingales, and D. Maestripieri (2009). Gender differences in financial risk aversion and career choices are affected by testosterone. *Proceedings of the National Academy of Sciences 106*(36), 15268–15273.

Schwartz, J., D. J. Steffensmeier, H. Zhong, and J. Ackerman (2009). Trends in the gender gap in violence: Reevaluating NCVS and other evidence. *Criminology 47*(2), 401–425.

Slemrod, J. (2019). Tax compliance and enforcement. *Journal of Economic Literature 57*(4), 904–954.

Steffensmeier, D. and E. Allan (1996). Gender and crime: Toward a gendered theory of female offending. *Annual Review of Sociology 22*, 459–487.

Swamy, A., S. Knack, Y. Lee, and O. Azfar (2001). Gender and corruption. *Journal of Development Economics 64*(1), 25–55.

Thakral, N. and L. T. Tô (2021). Daily labor supply and adaptive reference points. *American Economic Review 111*(8), 2417–2443.

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