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

DISCUSSION PAPER SERIES

IZA DP No. 15599

Mystery Shopping as a Strategic Management Practice in Multi-Site Firms

Sidney T. Block
Guido Friebel
Matthias Heinz
Nikolay Zubanov

SEPTEMBER 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15599

Mystery Shopping as a Strategic Management Practice in Multi-Site Firms

Sidney T. Block
University of Cologne

Matthias Heinz
University of Cologne

Guido Friebel
Goethe University Frankfurt and IZA

Nikolay Zubanov
University of Konstanz and IZA

SEPTEMBER 2022

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

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

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

www.iza.org

ABSTRACT

Mystery Shopping as a Strategic Management Practice in Multi-Site Firms*

Anonymous and unannounced site inspections known as "Mystery Shopping" (MS) are common in multi-site service firms, but little is known about the strategic importance of this practice. We conceptualize MS as a monitoring tool firms use to implement the optimal allocation of site resources between sales- and service-related activities in the presence of cross-site reputation spillovers, which is to maximize sales while maintaining service standards. Consistent with this view, data from three retail chains reveal (i) low variation in MS scores, (ii) little correlation of MS scores with sales, and iii) high correlation of sites' MS scores with the likelihood of their supervisors receiving incentive bonuses. Our findings are robust to different estimation specifications, and shed new light on a ubiquitous yet little-studied management practice.

JEL Classification: L10, M31, M41, M52

Keywords: mystery shopping, monitoring, reputation spillovers, incentives, service standards

Corresponding author:
Nikolay Zubanov
Department of Economics
University of Konstanz
Universitätsstrasse 10
78464 Konstanz
Germany
E-mail: nick.zubanov@uni-konstanz.de

* We appreciate valuable comments by Bernd Skiera, Christian Leuz, Martin Artz, Thomas Ruchti (discussant), Timo Vogelsang, Xiao Yu Wang (discussant), and seminar participants at ACMAR 2021, IOEA 2022, University of Cologne, and Goethe University Frankfurt. Jannis Heck, Isabel Münch, and Masha Reimers provided excellent research assistance. The project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1–390838866. Sidney T. Block would like to thank the Jürgen Manchot Stiftung for financial support. We did not receive funding from the study firms, and no co-author had a financial relationship with the study firms. We obtained approval by the study firms' work councils, and the study firms agreed to provide us with the data we use for research purposes. All opinions and errors are our own.

1 Introduction

Multi-site firms face a dual task of adapting to local conditions ("local responsiveness") while maintaining uniform service standards across sites to protect the brand value ("chain uniformity") (Cameron 1986; Bradach 1997). Much of the strategy and related literature discusses how these potentially conflicting goals can be reached through optimal delegation of decision rights from the central management to local sites, using franchise vs. own-operated firms as an exemplar setting (Brickley and Dark 1987; Shane 1996; Kaufmann and Eroglu 1999; Yin and Zajac 2004; Meiseberg 2013; Sorenson and Sørensen 2001; Lafontaine and Shaw 2005; Barthélemy 2008). Yet several important questions remain open.

In this study, we shed light on a fundamental problem both chains and franchise firms are faced with. On the site level, i.e. retail stores and restaurants, managers and workers allocate attention and effort to reach their preferred balance between the benefits of adapting to local circumstances (in order to increase sales) versus reaching the service standards (to maintain brand value). Chains and franchise firms are likely to care more about the brand value then their stores or restaurants. They, hence, need to influence the resource allocation problem on the site level through a number of instruments. Incentives or ownership alone, which are mostly treated in the literature on multi-site firms, may not suffice.

We argue that firms use a particular organizational practice, "Mystery Shopping" (MS), to achieve their goal of influencing the resource allocation process in retail stores and restaurants. MS is a common monitoring practice in multi-site firms which involves anonymous and unannounced site inspections by trained raters who evaluate pre-defined aspects of service delivery while acting as ordinary visitors (Finn and Kayandé 1999; Section 2.1 provides detail on the use of MS by firms). Based on data from MSPA (2018), one can estimate that, globally, there are dozens of millions of MS visits per year in numerous large industries, such as retail, finance, transportation and hospitality (Wilson 1998*b*; Beck and Miao 2003).

In investigating the use of MS as a tool to steer the resource allocation process of sites, we believe to add a multi-layered perspective to the problem of resource allocation. It has been noted before that the resource allocation literature is mainly concerned to understand how in multi-unit firms, top management allocates financial resources to divisions (Maritan and Lee 2017). Some recent work zooms in on non-financial resources, among others managerial attention (Lovallo and Sibony 2018). Agency relationships are part of the literature (Gibbons and Roberts 2013; Friebel and Raith 2010) and a source for inefficiencies that resource allocation process and strategies must overcome.

We here suggest an analysis in which top management uses a mix of output-contingent bonuses and action control through MS in the retail and restaurant sectors to find the right balance between actions that benefit a given site and actions that benefit the firm as a total, in particular its reputation. We spell out the agency relation between the center and the sites, show how MS may help to influence the resource allocation problem on the site-level, and present correlational evidence from three study firms in support of this interpretation.

Linking our results to the relevant literatures (Section 5), we see the key contribution of our study in conceptualizing and documenting empirical evidence on the use of MS as a strategic management practice used to address one of the strategic agency problems faced by multi-site firms – resource misallocation by individual sites – through monitoring and rewarding compliance with service standards. Maintaining service standards has been listed among the practical uses of MS in expert surveys by Wilson (1998*a*,*b*, 2001) and Beck and Miao (2003), but without linking it to strategic agency issues or other practices such as incentives.

2 Theory development

We begin this section by presenting MS as a strategic management practice in the context of the existing literature and identifying the research gap we aim to fill. We then concentrate on the agency problem focal to this study: misallocation of resources between sales- and service-related tasks by individual sites of a multi-site firm. We next outline the firm's strategy to deal with this problem, discuss the role of MS in implementing this strategy, and present three testable hypotheses.

2.1 Mystery shopping as a strategic management practice: Related literature and the research gap

Drawing on interviews with strategy scholars and text analysis of studies in major management journals, Nag, Hambrick and Chen (2007, p. 942) define strategic management as the field of research that "deals with the major intended and emergent initiatives taken by general managers on behalf of owners involving utilization of resources to enhance the performance of firms in their external environments." MS as a practice fits the above definition of "strategic". It is a major initiative, sponsored by firms' central management and involving significant resources.[1] MS focuses on monitoring compliance with service standards (Wilson 1998*a*,*b*, 2001), and is prevalent

[1] According to the Mystery Shopping Professional Association, firms spent about $2bn in 2016 on MS agencies employing mystery shoppers globally (MSPA 2018), a figure not including the costs of in-house MS programs and the costs of handling MS data for firms' internal purposes. Interviews with our study firms' representatives suggest that each firm spends about 25 to 50 Euros per MS visit. Hence, one can

in environments where maintaining these standards is especially important, namely, multi-site service firms where service interactions with customers is the core economic activity and brand reputation is a major value driver (Kidwell, Nygaard and Silkoset 2007; Gillis, Combs and Yin 2020).[2] MS as a service standard compliance practice was found to be effective in a variety of contexts: It is linked to a higher likelihood of crime registration and better handling of cases by police officers in India (Banerjee et al. 2021), less over-prescription of drugs by Chinese physicians (Cheo et al. 2020), and better service ratings and customer count in US restaurants (Latham, Ford and Tzabbar 2012).

Strategic management initiatives being "taken by general managers on behalf of owners", as per definition above, links strategy with "agency problems", or conflicts of interest between different parts of the firm. Particularly related to MS is the agency problem of sub-optimal effort allocation between sales- and service-related tasks by individual sites, which we describe in detail in the next section.

In the absence of perfect action controls, a solution to the above agency problem would be to provide incentives based on service performance in addition to sales-based incentives. We know from existing studies that some firms do just this (e.g., Gibbs et al. 2004; Campbell 2008; Bouwens and Kroos 2017), and so do all three of our study firms. Firms use service performance metrics from customer surveys as well as MS (Jacob, Schiffino and Biard 2018), but MS-based metrics have been found to be more reliable than those from customer surveys, presumably because MS raters are trained to monitor certain specific actions by the site personnel while customers are only able to provide their overall, vague feeling about a service encounter (Finn and Kayandé 1999; Finn 2001, 2007). Yet, little is known about whether and how the practices of MS and incentives are linked within firms. Existing literature only briefly notes that firms might use MS in providing employee incentives (Erstad 1998; Wilson 1998*b*; Blessing and Natter 2019) without theoretical elaboration or empirical evidence.

estimate that, globally, around 40 to 80 million MS visits are carried out each year. In addition to MS, there are statutory inspections and checks carried out by authorities rather than firms themselves. These checks, however, focus more on preventing health risks rather than on assuring customer service, and are therefore beyond the scope of this study.

[2]For example, Beck and Miao (2003)'s survey of U.S. hospitality establishments finds that 86.6% of hotels use MS programs.

Theory and evidence on MS as a strategic management practice are both needed to bring this important topic closer to the strategy research. Both are currently missing, and the strategy literature on MS is non-existing.[3] Our study aims to bridge this research gap by linking the practice of MS to the specific agency problem multi-site firms face, and by documenting empirical evidence that supports this link.

2.2 The agency problem of resource misallocation in a multi-site firm

We first present a verbal summary of the agency problem we focus on in this study, delineating it from other agency problems in multi-site firms, and then proceed to its more rigorous mathematical formalization which also helps generate testable hypotheses.

2.2.1 A non-technical summary

Our theoretical setting builds on Brickley and Dark (1987) and is as follows. There is a firm managing a network of autonomous and geographically dispersed sites. The firm aims to maximize the total of the revenue streams from its sites net of the total costs, but the individual sites maximize their income net of their costs. Each site has a resource budget which it allocates between the two groups of tasks which we label as "sales" and "service". Sales tasks generate sales at individual sites, but have little consequence beyond, for example, routine technological operations or processing sales transactions. On the other hand, service tasks such as cleaning, product presentation or staff-customers interactions contribute to the reputation of the firm as a whole, and thus affect sales beyond focal sites.

One agency problem immediately following from this setting is effort under-provision by sites ("shirking and perquisite-taking" in Brickley and Dark (1987)): because effort is costly, site workers on a fixed salary may choose to provide less effort than the firm would find optimal, resulting in site resources being wasted rather than spent on sales or service. Solutions have been offered to this well-known problem, including franchising or other schemes that link a site's reward with its sales (Jensen and Meckling 1976;

[3]Searching through major strategy and general-interest management journals (*Strategic Management Journal, Academy of Management Journal, Academy of Management Review, Administrative Science Quarterly, Management Science*, and *Journal of Management*) for articles that have the terms "mystery shop*" or synonyms ("mystery cust*", "secret shop*", "secret cust*", and "decoy visit") in the title or abstract, we found *no* single article that has any of our search terms in the title, and only two articles that have our search terms in the abstract. Both studies used MS data for research purposes not related to MS as a management practice. Using the same key terms, we established the business and management journals most likely to publish studies on MS, which are: International Journal of Marketing Research (2/3), Journal of Hospitality Marketing & Management (3/3), Journal of Quality Assurance in Hospitality and Tourism (4/8), Journal of Retailing (2/2), Journal of Service Research (1/2). The numbers in parentheses refer to the number of matches with the search terms in the title / abstract. Beyond business management journals, studies on MS are also published in healthcare and library management journals.

Fama and Jensen 1983). We assume from the beginning that sites receive sales-based incentives, so that no site resources are spent unproductively. Indeed, sales-based incentives is a usual practice in multi-site firms (Nyberg et al. 2018), and all our study firms use them.

The agency problem we focus on is *misallocation of resources between sales and service tasks by individual sites*, often referred to as "free riding by the franchisee" in the franchising literature (Brickley and Dark 1987; Michael 2000; Jin and Leslie 2009; Helm and Salminen 2010). The cause of this problem is *reputation spillovers*: service received at one site may affect sales at other sites under the same brand. For example, a customer dissatisfied with the service received at one site may be less likely to visit not only that specific site but also other sites operated under the same brand, increasing the total sales loss from bad service beyond the level felt by individual sites. Reputation spillovers make the marginal product of service effort bigger for the firm than for individual sites. As a result, the service effort chosen by individual sites that receive sales-based incentives is too low from the firm's perspective.

Service effort provision by individual sites being too low is not the only issue. When site resource budgets differ (they do, empirically), bigger sites with larger resource budgets would spend more effort on both sales and service than will smaller sites, since both sales and service efforts contribute to their sales. However, reputation spillovers lead to sub-standard service at one site harming the entire firm's reputation more than excellent service at another site can improve it. As a result, individual sites' service effort choices are not only too low from the firm's perspective, but also they vary too much across sites depending on their resource endowments.[4]

Service experienced at one site affecting other sites through reputational spillovers creates an externality that cannot be repaired by incentivizing site sales alone. Additional incentives, based on service metrics rather than sales, are necessary to raise service efforts.[5] In practice, such incentives often come in the form of bonuses for meeting firm-imposed *service standards*, or maluses for failure to meet them (Bradach 1995; Ingram and Baum 1997; Bradach 1998; Sorenson and Sørensen 2001). Offered this way, service-based incentives are meant not only to raise service efforts but also to homogenize them across sites, thus bringing sites' effort allocation between sales and service tasks closer to the firm's optimum. MS is instrumental in enforcing service standards through monitoring and incentives based on MS reports.

[4] Brickley and Dark (1987) discuss two other agency problems faced by multi-site firms, inefficient risk-bearing by franchise holders and quasi-rent appropriation through hold-up behavior by either franchiser or franchisee, but these issues are more specifically relevant to franchise firms than firm-owned multi-site operations (we study both) and less central to our research questions.

[5] As an analogy to cross-site service externalities and additional incentives focused on service, it may be helpful to think of public health externalities caused by private car use and the resulting surcharges on fuel that are meant to incentivize "greener" mobility choices.

2.2.2 A formal model

To support the above intuitions, we now develop a simple formalization of the agency problem of resource misallocation by individual sites. Let site i in a network of size n be endowed with a resource budget b_i (say, working time) of which it spends a quantity s_i on service and the remainder $p_i = b_i - s_i$ on sales tasks as defined earlier. Both sales and service efforts affect the focal site's sales through the production function $g(p_i, s_i)$ with a positive and diminishing rate of return with respect to both effort types: $\frac{\partial g}{\partial x} > 0, \frac{\partial^2 g}{\partial x^2} < 0, x \in \{s_i, p_i\}$. The sites receive sales-based incentives in the form of share $\alpha \in (0, 1]$ of sales.

In addition to contributing to site i's sales through the production function $g(p_i, s_i)$, service effort s_i affects the firm's reputation which we model with the reputation factor $r(s_1, s_2, ..., s_i, ..., s_n) > 0$ that, thanks to reputation spillovers, depends on the service efforts at all sites. As is usual, we assume that reputation responds to service effort at a positive but diminishing rate ($\frac{\partial r}{\partial s_i} > 0, \frac{\partial^2 r}{\partial s_i^2} < 0$), and that service efforts at different sites are complementary to each other in creating the firm's overall reputation ($\frac{\partial^2 r}{\partial s_i \partial s_j} > 0$). That is, better service at one site makes service effort at another site more effective, and, conversely, good service at one site contributes less to the overall reputation if service at other sites is bad.

The firm maximizes the long-term value generated through repeat custom enabled by reputation for service earned by all of its sites together. Specifically in our model, we capture the long-term value generated at site i as the product of its sales and the reputation factor[6]: $LV_i = r(s_1, s_2, ..., s_i, ..., s_n) \cdot g(p_i, s_i)$. Hence the firm's optimization problem,

$$\max_{s_i} \sum_{i=1}^{n} LV_i = r(s_1, s_2, ..., s_i, ..., s_n) \cdot \sum_{i=1}^{n} g(p_i, s_i),$$

the solution to which in terms of the service effort at each site (s_i^*) is determined by the following first-order condition

$$\frac{\partial \ln(r(..., s_i^*, ...))}{\partial s_i} = \left[\frac{\partial \ln(g(p_i^*, s_i^*))}{\partial p_i} - \frac{\partial \ln(g(p_i^*, s_i^*))}{\partial s_i} \right] \cdot \phi_i, \tag{1}$$

where $\phi_i = \frac{g(p_i, s_i)}{\sum_{i=1}^{n} g(p_i, s_i)} < 1$ is the share of site i's sales in the total, and the optimal sales effort is $p_i^* = b_i - s_i^*$. Detailed derivations of this and the following analytical results is available in Appendix C.

[6] For an intuition, think of the long-term value as the infinite stream of one-period sales, g_t, discounted by a factor δ that depends on the probability of a repeat custom, which in turn depends on service: $\sum_{t=0}^{\infty} g_t \delta(s)^t = \tilde{g} \cdot r(s)$, where $r(s) = \frac{1}{1-\delta(s)}$ and $\tilde{g}$ is some weighed average of sales in all time periods. Note that in a more general version of our model the optimal service level may vary with time, reflecting the build-up and wearing-out of reputation over time. However, in this paper we abstract from these dynamics for simplicity and also because the time span of our data is limited.

Individual sites, however, maximize their income from sales-based incentives, not the firm's total:

$$\max_{s_i} r(s_1, s_2, ..., s_i, ..., s_n) \cdot \alpha \cdot g(p_i, s_i).$$

The optimal service effort s_i^{**} that solves the above problem is given by

$$\frac{\partial \ln(r(..., s_i^{**}, ...))}{\partial s_i} = \left[\frac{\partial \ln(g(p_i^{**}, s_i^{**}))}{\partial p_i} - \frac{\partial \ln(g(p_i^{**}, s_i^{**}))}{\partial s_i} \right] \cdot 1, \tag{2}$$

where $p_i^{**} = b_i - s_i^{**}$ is the optimal sales effort.

Comparing the optimal service effort levels from the firm's and sites' perspectives, s_i^* in (1) and s_i^{**} in (2), respectively, demonstrates the first aspect of the agency problem outlined above: *sites choose lower service effort levels than what the firm would find optimal.* Recalling that $\frac{\partial^2 r}{\partial s_i^2} < 0$ (diminishing returns to service effort) and $\phi_i < 1$ (a single site's share in the total sales is < 1), it follows that $s_i^{**} < s_i^*$.[7]

The second aspect of the agency problem, that *sites choose more varying service effort levels than what the firm would prefer*, can be seen by comparing cross-site variation in the firm- (s_i^*) and site-optimal (s_i^{**}) service effort levels. While sites with a larger budget b_i should optimally provide higher service levels ($\frac{ds_i}{db_i} > 0$ for both firm- and site-preferred s_i), the firm-preferred service level s_i^* in (1) is less sensitive to the site's budget b_i than is the site-preferred s_i^* in (2):

$$\frac{ds_i^*}{db_i} < \frac{ds_i^{**}}{db_i} \tag{3}$$

Therefore, the firm would like not only higher but also less varying service levels than what individual sites would choose.

2.3 The firm's strategy to address the resource misallocation problem

In this section we outline a strategy to deal with the site resource misallocation problem through imposing, monitoring and rewarding sites' compliance with service standards, and present testable hypotheses that would hold if this strategy is implemented.

[7] This prediction of our model does not imply that chains should provide better service than independent establishments. Independent establishments are not part of a chain, so reputation spillovers, which cause the difference between the firm- and site-preferred service levels in our model, are presumably weaker. Besides, chains and independent establishments are likely to differ in technology, target customer groups and other characteristics that affect the productivity of service effort, making the difference between the two business models even more nuanced.

2.3.1 Service standards

Even though the theoretically optimal service levels vary across sites (equation 3), the firm may prefer imposing uniform service levels, or *service standards*, instead of trying to implement site-specific optimal service levels. There are several reasons for doing so: ease of administration; savings from uniform staff training; customer preferences for standardization when quality is ex-ante uncertain; large network size which makes individual sites relatively small (as can be seen from the derivations in Appendix C, low ϕ_i limits the variation of optimal service levels with site endowments). Besides, optimal service variability is further suppressed by cross-site service effort complementarities in the reputation factor $r(...)$. Complementarity formally means that the marginal effect on the firm's sales of service effort in a given site increases with the service efforts in the other sites: $\frac{\partial^2 r}{\partial s_i \partial s_j} > 0$ (Milgrom and Roberts, 1990), which implies that good service at one site has a larger effect on the firm's reputation when service at other sites is also good, and is ineffective when service is bad elsewhere. Complementarity is thus consistent with reputation spillovers we assumed earlier.

To show the effect of service effort complementarity on the variation of the optimal service efforts across sites, consider a particular (but still rather flexible) "constant elasticity of substitution" (CES) specification for the reputation factor: $r(s_1, ..., s_i, ..., s_n) = a \cdot \left(\sum_{i=1}^{n} s_i^{\rho}\right)^{\frac{1}{\rho}}$, where $a > 0$ is a scale constant and ρ is the complementarity parameter: $\rho < 1$ implies complementarity (which we assume), and $\rho > 1$ implies substitutability.[8] The variation of the optimal service effort with the site's resource budget, $\frac{ds_i^*}{db_i}$, increases with ρ, implying lower variability in the optimal service levels across sites under stronger complementarity (lower ρ). In the extreme case of $\rho \to -\infty$, $r(s_1, ..., s_i, ..., s_n) = a \cdot \left(\sum_{i=1}^{n} s_i^{\rho}\right)^{\frac{1}{\rho}}$ converges to $a \cdot \min(s_1, ..., s_i, ..., s_n)$. Put differently, complementarities are so strong that the worst site service determines the firm's overall reputation, and the optimal service level is the same across the sites regardless of resource budget size, that is,

$$\lim_{\rho \to -\infty} \frac{ds_i^*}{db_i} = 0 \Rightarrow \lim_{\rho \to -\infty} \frac{dp_i^*}{db_i} = \lim_{\rho \to -\infty} \frac{d[b_i - s_i^*]}{db_i} = 1 \tag{4}$$

[8] By definition, $\frac{\partial^2 r}{\partial s_i \partial s_j} > 0$ means complementarity. Taking the cross-derivative of $r(s_1, ..., s_i, ..., s_n) = a \cdot \left(\sum_{i=1}^{n} s_i^{\rho}\right)^{\frac{1}{\rho}}$ with respect to any pair s_i, s_j, we obtain $\text{sign}\left[\frac{\partial^2 r}{\partial s_i \partial s_j}\right] = \text{sign}(1 - \rho)$; so $\rho < 1$ means complementarity. CES aggregators have been used in economics research to formalize complementarity between inputs in creating aggregate output (e.g., individual members' contributions in the team's output) (Iranzo, Schivardi and Tosetti 2008; Friebel et al. 2017, 2021).

Thus, under strong complementarities and economies from uniform service provision, ensuring that sites *maintain service standards and spend the rest of their resources on generating sales* may be the firm's most practicable strategy of dealing with effort misallocation by individual sites.

2.3.2 Monitoring and rewarding compliance with service standards

Since service effort levels corresponding to the standards will not be freely chosen by individual sites, careful measurement of site service performance is required for incentivizing their compliance with the standards. MS is well-placed to evaluate site service performance. While sales are relatively easy to record, obtaining reliable service metrics is complicated by "noise" inherent in subjective impressions of service quality (Finn and Kayandé 1999; Blessing and Natter 2019), and possibly by "signal manipulation" through the actions of site employees to improve evaluated service performance during inspection (Makofske 2020). The practice of MS takes both these complications into account. To prevent signal manipulation, MS inspections are purposely unannounced and the mystery shoppers are instructed to behave like usual customers, and may even be rotated between sites not to look too familiar to site staff (they are in our study firms, see Section 3.2). To reduce noise, MS routines are standardized by training mystery shoppers to record specific, well-defined aspects of service delivery and rate them using well-defined scales (see appendix B for examples of MS checklists). MS routines' being unannounced to sites and following a rigorous, standardized script enhances the reliability of MS reports as compared to regular customer surveys (Finn 2001, 2007) and improves the suitability of MS scores as a performance metric for incentives purposes.[9]

Turning to rewards, supporting the firm's strategy of encouraging sites to commit a fixed amount of resources to maintain service standards and spend the rest to stimulate sales is an incentive scheme where the receipt of a bonus is linked to the MS score reaching a certain threshold corresponding to the service standard, and the bonus size grows monotonically with sales. This way the sites will have no incentive to economize on service effort when it is below-standard but also no incentive to excel in service when it is above-standard, directing the remaining resources to sales activities instead.

[9]Noisiness of performance metrics is an important consideration for their usability to support incentives: metrics with less noise can support stronger incentives (e.g., Baker 2002).

2.3.3 Testable hypotheses

To summarize our theory, reputation spillovers result in effort misallocation problem, whereby individual sites choose lower and more varying service effort levels than is optimal for the firm. A practical strategy of dealing with this problem is to impose service standards, monitor them with MS and reward compliance by offering bonuses whose receipt depends on meeting the standard and whose size depends on sales.

Assuming the firm succeeds in implementing the above strategy, the variation in the optimal sales effort across sites with different resource budgets together with the lack of variation in the optimal service effort (equation 4) will result in a low variance in service performance metrics relative to that of sales. Hence our first testable hypothesis:
H1. Sites will vary in MS scores less than in sales.
Maintaining service standards will also result in a low correlation between service and sales metrics, since optimal sales efforts change in lockstep with the resource budget, whereas service efforts are uniform. Hence our second hypothesis:
H2. There will be little correlation between sites' MS scores and sales.

The relationship between the receipt and size of the bonus and MS score or sales may be fuzzier than its theoretical prediction. Noise in performance metrics as well as additional information about site-specific circumstances available to the firm management call for a degree of discretion in awarding bonuses.[10] Rather than a mechanical rule converting performance metrics into cash, discretion in awarding bonuses implies a positive correlation between MS score and the likelihood of receiving a bonus, and a positive correlation between bonus size and sales. Hence our third hypothesis:
H3. A site's MS score will affect the likelihood of it receiving a bonus whose size will monotonically increase in sales.

3 Study background

In this section, we describe our three study firms. All firms are multi-site retailers. Although they differ in size, location and scope (see descriptive statistics in Table 1), their MS practices are similar to each other and to other firms in the same line of business.

[10] Baiman and Rajan (1995) and Rajan and Reichelstein (2006) present theoretical arguments for managerial discretion in bonus awards, and Gibbs et al. (2004), Bol et al. (2010), Bol, Hecht and Smith (2015), and Arnold and Tafkov (2019) document supporting empirical evidence.

3.1 The study firms

Firm 1 (Table 1, panel A) is a large bakery chain selling fresh bread, rolls, snacks, and drinks. Its 368 sites (bakeries) are located in big and mid-sized cities all over Germany, Austria and the Netherlands. The sites are franchised to independent operators who contractually must follow the same business strategy, with product assortment, visual presentation and site design stipulated in the franchise contract.

The average site in Firm 1 generates about 52k Euros worth of sales per month and employs just over eight workers (six in full-time equivalence) who have an employment contract with the franchise holder. The workers are busy with preparing pre-fabricated food and cutting fresh sandwiches, displaying the products neatly, cleaning the site, and providing customer service typical of a bakery. The employment contracts vary by site; workers receive close to minimum wage and some franchise holders pay small discretionary bonuses to their workers. Franchise holders pay a fixed franchise fee and a percentage of net sales to the firm, retaining the rest (net of the fixed franchise fee, their total pay is 93-95% of net sales).

Firm 2 (Table 1, panel B) is a smaller, regional bakery chain selling a range of products similar to Firm 1's. Firm 2's sites are located in a sub-urban region in central Germany spanning roughly 100 x 60 kilometers. During the period of observation (January 2012 - December 2014), the average site sold about 26k Euros worth of products per month. All Firm 2's 232 sites are operated by the firm itself. The sites have on average seven employees (four in full-time equivalence), including the site supervisor. All workers receive close to the minimum wage, and additionally a share of the team bonus for reaching sales targets.[11] Their tasks are similar to those of the workers in Firm 1.

Site supervisors receive a baseline fixed salary plus a discretionary bonus (8% of the total salary, on average). We learned from the interviews with Firm 2's management that bonuses are paid taking into account sites' performance with respect to their sales and personnel costs targets as well as MS scores, and tend to be set as a percentage of the baseline salary. Bonuses are paid monthly. It takes about two months to process performance records and decide on the bonus amounts; so performance results now will register in the bonus paid in two months' time.

[11] Employee bonuses conditional on sales were introduced experimentally in April-June 2014, and then the scheme was rolled out to the entire firm. When available, team bonuses were small, about 2% of the average worker's salary. However, note that we do not have MS data after May 2014. Data from firm 2 have been used in Friebel et al. (2017), Khashabi et al. (2021) and Friebel et al. (2021) to address questions unrelated to MS. The number of sites in this study (232) is larger than in those papers (193) because the other papers used data only from the sites operational as of the beginning of the treatments studied there, whereas in this study we include all sites ever operated by firm 2 during the observation period.

Table 1: Site characteristics

	Mean	St. dev.	Between st. dev.	Within st. dev.
Panel A: Sites in Firm 1 (franchise, bakeries)				
Monthly sales (€)	51,942	35,517	33,665	6,037
Daily sales (€)	1,867	1,267	1,093	489
Monthly hours worked	877	651	599	219
Monthly headcount	8.45	4.66	4.23	1.89
Franchisees' monthly salary	93-95% of total sales			
Observation period	24 months: Feb. 2016 - Jan. 2018			
Number of sites	368			
Panel B: Sites in Firm 2 (integrated, bakeries)				
Monthly sales (€)	26,408	12,926	12,263	3,122
Monthly hours worked	705	335	320	106
Monthly headcount	6.94	3.02	2.79	1.19
Supervisors' monthly salary (€)	1,880	237	239	16
Supervisors' monthly bonus (% of monthly salary)	7.96	8.45	4.96	7.25
Observation period	36 months: Jan. 2012 - Dec. 2014			
Number of sites	232			
Panel C: Sites in Firm 3 (integrated, supermarkets)				
Monthly sales (€)	213,298	156,651	152,409	35,104
Monthly hours worked	3,355	2,415	2,383	352
Monthly headcount	24.72	17.91	17.68	2.59
Supervisors' monthly salary (€)	901	195	110	162
Supervisors' quarterly bonus (% of quarterly salary)	5.02	7.19	3.12	6.48
Observation period	41 months: Jan. 2014 - May 2017			
Number of sites	241			

Notes: The daily sales are only available for firm 1. The number of sites refers to the number of sites for which we have mystery shopping data. The observation period refers to the number of months for which we have either mystery shopping or sales data. We report the mean, and the overall, between- and within-site standard deviations of mystery shopping scores. The square of the between-site standard deviation and the square of the within-site standard deviation do not sum up exactly to square of the total standard deviation because some sites contribute more site-month observations to our data set than others.

Firm 3 (Table 1, panel C) is a leading supermarket chain in an Eastern European Union (EU) country comprising 241 own-operated sites (grocery stores) spread over the entire country. The sites sell a wide range of grocery products, some of the larger sites having in-house bakeries, fishmongers and other specialized departments. The average site sells 213k Euros worth of goods and employs just under 25 workers. The largest employee group, 82% of the workforce, are general site assistants ("cashiers") who receive a close to minimum wage plus a performance-related bonus (about 5% of the total salary, on average). Their tasks include shelving, operating cash tills and keeping the site clean during the day. The other worker groups are "specialists" and unit managers running site departments (e.g., the in-house bakery), where available.[12]

[12]Data from firm 3 have been used in Friebel, Heinz and Zubanov (2022) and Friebel et al. (2022) to address questions unrelated to MS.

Firm 3's site supervisors receive a baseline salary plus a discretionary bonus (5% of the total salary, on average). As we were told by Firm 3's management, the bonus is based on similar performance indicators as in Firm 2: sales vs. targets, and MS scores, but it is not a convention there to set it as a percentage of the baseline salary. Firm 3 pays bonuses quarterly; so performance results in a given quarter would register in the next quarter's bonus.

The site supervisors in all three firms spend most of their time on-site, helping in, overseeing the logistics and doing administrative tasks such as personnel planning and running the accounts. Most importantly, site supervisors also manage employees' allocation of efforts between tasks. The relatively small size of the sites in Firms 1 and 2 and a high degree of personal involvement of site supervisors enable close supervision and oversight. For all firms, we abstract from within-site agency issues and focus on those between the sites and central management, as described in Section 2.2.

All firms monitor as well as incentivize their site managers. Predictably, monitoring is more intense and incentives are weaker in the integrated, non-franchise Firms 2 and 3 than in the franchise Firm 1. Indeed, as we learned from site supervisor surveys in Firms 2 and 3, their regional managers (the next hierarchy level) regularly visit sites and monitor site activities and performance results. Contrarily, regional managers in Firm 1 do not personally visit or otherwise inspect sites but only communicate between the firm and the franchise holders. In addition to sales-based incentives, all firms pay MS-based incentives to site supervisors (detail in Section 4.3).

3.2 Mystery shopping practices in the study firms

All of our study firms hire external agencies to provide MS, who send their trained mystery shoppers to visit the firms' sites and rate service quality, which is a common practice (Wilson 1998*b*). Sites are visited fairly frequently: an average site is 97% likely to be visited in an average month in Firm 1, 81% in Firm 2, and 94% in Firm 3 (Table 2). A large number of mystery shoppers are employed: 381 in firm 1 and 355 in Firm 2 (no such data available for Firm 3). To preserve anonymity their anonymity from site personnel, mystery shoppers are rotated between sites: In our observation period, the average mystery shopper visits 5.5 (9.2) different sites during their total of 13.0 (11.3) visits in Firm 1 (Firm 2) (see Table 2 for descriptive statistics on MS visits). The large number of mystery shoppers relative to the number of sites is typical of mystery shopping as an occupation. According to The U.S. Bureau of Labor Statistics' Career Outlook, mystery shopper is a flexible, part-time job, and mystery shoppers register with many different providers (Torpey 2016). In our our study firms, a MS visit costs about 25 to 50 Euros, an amount comparable to Finn and Kayandé (1999)'s estimate of 60 U.S. Dollars per visit.

Table 2: Mystery shopping statistics

	Firm 1	Firm 2	Firm 3
Number of mystery shopping visits	8,115	4,965	8,654
Number of sites	368	232	241
Number of mystery shoppers	381	355	-
Number of observed months	24	36	41
Months with mystery shopping visit (%)	97.2	80.8	94.1

Notes: The number of sites refers to the sites for which we have mystery shopping data. The number of observed months refers to the number of months for which we have either mystery shopping or sales data. Mystery shoppers identifiers are not available for firm 3. The fraction of months with a mystery shopping visit is the fraction of months the average site receives at least one visit per month.

Mystery shoppers rate their experiences using a standard evaluation form that lists the positions to be evaluated (templates for Firms 1 and 2 are in Appendix B; we have no information on the template used by Firm 3). In Firms 1 and 2, the multiple positions evaluated by mystery shoppers can be grouped into three categories. Those categories correspond closely to the tasks of the site employees (Section 3). The positions in category *personnel* instruct mystery shoppers to record whether employees are present at the point of sales, are friendly and welcoming, have clear speech, wear clean work clothes, offer additional products, ask for a loyalty card (where applicable), and say "thanks" and "goodbye". The category *product* contains positions related to product availability, presentation, taste and freshness. The category *cleanliness* evaluates how clean and tidy the counter, seating area, plates, floor, coffee machine, coffee area and fridges are. To reduce subjectivity, many positions are phrased in objective terms (e.g., "was there at least one employee present when you entered the site?"), and where this is not possible (e.g., cleanliness) examples of "clean" and "not clean" are provided in the memo sent to mystery shoppers (we have seen the memos but cannot provide them for reasons of anonymity). The MS evaluation forms in firms 1 and 2 are similar in structure and content to those used by other service firms selling food (e.g. retail or restaurants).[13]

While the study firms track the MS scores by category, they pay incentives to site supervisors based on the *aggregate MS score*, calculated as the weighted average of the category scores standardized to vary between 0 (the lowest) and 100 (the highest possible score). The *personnel* category enters with the highest weight, followed by *product*

[13]We received access to MS questionnaires used by the retail chain featured in Manthei, Sliwka and Vogelsang (2021), akin to our firm 3, which include very similar positions to those in Firms 1 and 2's forms. In addition, the fast food retailer in Campbell (2008)'s study uses criteria for promotion some of which are alike to the categories our study firms survey with MS. In the 2000s, when Campbell (2008)'s study was carried out, McDonald's also surveyed its sites along the personnel, product and cleanliness dimensions (Wall Street Journal 2001).

and *cleanliness*.[14] MS-based incentives vary by firm: they are rules-based in the franchise Firm 1 and are discretionary in the integrated Firms 2 and 3. Section 4.3 provides further detail on the relationship between MS scores and site supervisor incentives. Site supervisors are aware, from communications with the central management, that the aggregate MS score matters for their bonus and how it is calculated in their firms. MS-based incentives seem to be effective in influencing sites' actions.[15]

4 Results

In this section, we present empirical results corresponding to our study hypotheses, and some auxiliary findings.

4.1 Variation in mystery shopping scores versus site sales (hypothesis H1)

Table 3 reports descriptive statistics of the MS scores observed in our study firms. The average aggregate MS scores are high: 88% in Firm 1, 96% in Firm 2, and 97% in Firm 3. The distribution of MS scores is compressed: the standard deviations of the aggregate MS score are 5-10% of the average, depending on the firm. Most of the variation in MS scores comes from within sites. The between-site variation is minor, 8-11% of the total (see Tables 6-8 in Appendix A for details on the analysis of variance). The mystery shopper "fixed effects" are an important variance component, accounting for about 20% of the total variance, or twice as much as site-level factors, reflecting considerable subjectivity in MS assessments.

The pattern of the variation in sales is very different from that in MS scores. The variation in sales is much larger: the standard deviation of sales is about or more than half of the mean in all firms (recall Table 1), five to ten times that of MS scores. Levene's variance comparison test for the aggregate MS score and sales, both transformed in percentage deviations from the mean to align scales, strongly rejects the null hypothesis of equal variances in sales and MS scores (p-value < 0.001), despite considerable noise in MS scores owing to mystery shopper fixed effects. Finding a significantly lower variance in MS scores than in sales supports our hypothesis H1: sites will differ in MS scores less than in sales.

[14] Specifically, Firm 1 chposes weights of 44.7% for *personnel*, 20.3% for *product*, and 35% for *cleanliness*. In Firm 2, the weights are 52.2%, 31.4% and 16.4% for the respective categories. The above weights are averages over the observation periods, as there have been minor scoring changes.

[15] A change in MS scoring in firm 1, happened in January 2017, provides some evidence. The change concerned asking the "eat in or take away?" question, which is important for the correct calculation of the sales tax in Germany (the tax is higher for eating in). Before the change, asking the question counted for a modest 2% of the aggregate MS score, and it was asked in 37% of cases. After a fine of 50 Euros was imposed for not asking the question, it was asked 65% of the time.

Table 3: Descriptive statistics on mystery shopping scores

	Mean	St. Dev.	Between St. Dev.	Within St. Dev.
Panel A: Firm 1				
Mystery shopping score	87.81	9.89	4.36	9.07
Mystery shopping cleanliness score	89.77	15.10	5.50	14.14
Mystery shopping product score	87.70	16.89	7.61	15.59
Mystery shopping personnel score	86.36	11.53	4.50	10.75
Panel B: Firm 2				
Mystery shopping score	96.22	5.09	1.81	4.87
Mystery shopping cleanliness score	98.51	6.31	2.14	6.13
Mystery shopping product score	97.48	5.68	1.57	5.50
Mystery shopping personnel score	94.84	7.58	2.81	7.24
Panel C: Firm 3				
Mystery shopping score	97.04	3.97	1.09	3.82

Notes: This table reports descriptive statistics for the aggregate mystery shopping score and its components, where available. The observations are at the site-month level. If there are more than one mystery shopping visits in a site-month, we average the mystery shopping scores. In firm 1, there are 387 site-months with two and 5 site-months with three mystery shopping visits. In firm 2, there are 29 and 6, respectively. In firm 3, no site is visited more than once a month. We report the mean, and the overall, between- and within-site standard deviations of mystery shopping scores. The between- and within-site variances in mystery shopping scores do not exactly add up to the overall variance because some sites are observed longer than others.

Contrary to MS scores that vary mostly within-site, the variation in sales is predominantly between-site. This difference in the variance structure of sales and MS scores is also consistent with our model that predicts that sites should optimally spend a certain amount of their resources to maintain (close to) uniform service standards, and use the rest of the available resources to generate sales. Compressed MS scores, especially between sites, is a manifestation of sites' maintaining service standards, while varying sales, especially between sites, reflects differences in resource endowments that are more pronounced between sites (e.g., location or management talent) than within. To summarize, in line with the optimal firm strategy from our model, we find that MS scores are uniformly high across sites, and that site performance differs more in terms of MS performance than in terms of sales performance.

4.2 Mystery shopping scores and site sales (hypothesis H2)

To estimate the correlation between MS scores and sales, we run several linear regression specifications with varying controls. The dependent variables in our regressions are the standardized (i.e. original score minus mean divided by standard deviation) aggregate MS score and its components (components are available in Firms 1 and 2). The key regressor is log sales on the day of the visit (Firm 1, for which we have daily sales) or log

monthly sales (Firms 2 and 3, for which we have monthly sales). As controls we include the time of the MS visit (morning, afternoon, evening), the month, mystery shopper (where available) and site fixed effects. The detailed regression output is shown in Tables 9, 10 and 11 in Appendix A.

Figure 1 plots changes in the standardized MS scores associated with a 10% increase in sales produced by different regression specifications. The results are quite comparable across the firms and specifications. A massive 10% increase in sales is linked to zero or only marginal changes in the aggregate MS scores, never more than 1% of their standard deviations. The same holds true when we control for potentially non-randomly missing monthly MS visits at sites in Firm 2, where missing monthly MS visits are relatively frequent (19% of site-months, Table 2), by means of estimating a Heckman selection model which results are reported in Table 12 in Appendix A. The results are similar and also show that whether a MS visit is missing or not is unrelated to site characteristics. The finding of little to no correlation between MS scores and sales is in line with our hypothesis H2.[16]

One explanation for the low correlation between sales and MS scores that would be alternative to that provided by our model is lack of statistical power to detect a significant correlation. Power does not seem to be an issue because the very small correlations we find are in fact statistically significant in many specification we run (Tables 9, 10 and 11 in Appendix A), so we find precise zero correlations rather than large but imprecisely estimated ones. Another explanation is the noise in MS scores that biases the estimated correlation with sales towards zero. While reliability of MS scores is clearly an issue (recall the analysis of variance in Tables 6 and 7, and Section 4.1), our results remain largely unchanged when we control for the likely sources of noise in MS scores, namely, site, month, time of the visit and mystery shopper fixed effects. Relatedly, regressing a MS score taken on a particular day on sales aggregated over a month may also bring in measurement error. However, we do not see a large difference between the estimates from Firm 1, where we have sales data on the day of the visit, and from Firm 2 where only monthly data are available.

Another explanation we pursued is reverse causality between MS scores and sales. Our specifications so far, where MS scores are modelled as endogenous to demand captured in sales, have been driven by the mechanism we model, which has to do with resource allocation within sites. However, it may be that service quality captured by MS scores positively affects future sales (e.g., Heskett et al. 1994; Anderson and Mittal 2000; Blessing and Natter 2019), thus weakening the correlations we plotted in Figure

[16]Figure 1 also shows that sales correlate positively with the MS-product score (medium gray), and negatively with the MS-cleanliness score (dark gray). These correlation, though statistically significant, are economically small and thus do not contradict our hypothesis H2. One could try to rationalize these correlations (e.g., sites may give cleanliness lower priority when demand is high), but given their size and lack of robustness to specification in the case of Firm 2, we refrain from over-interpreting them.

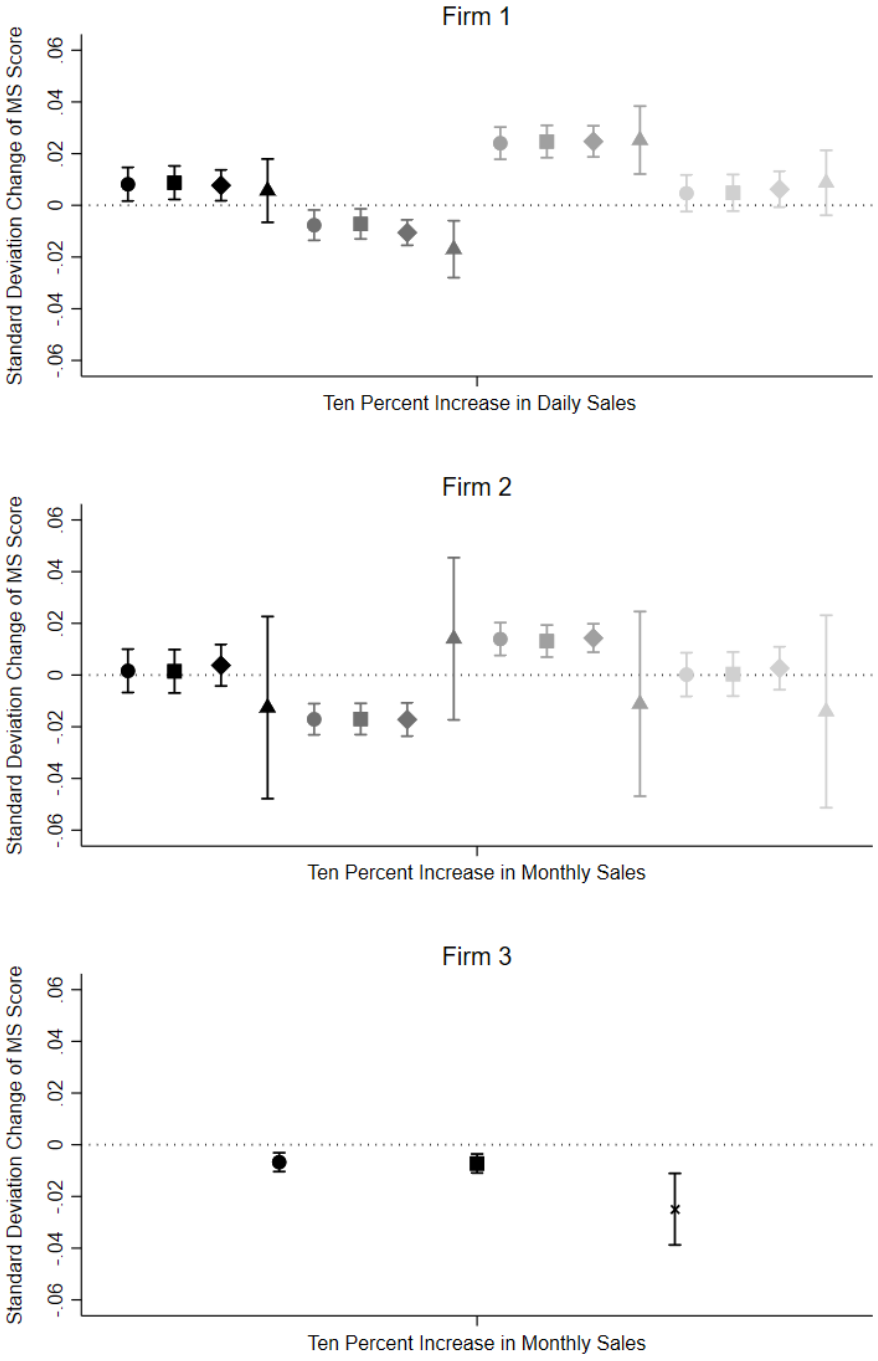

Figure 1: Correlation between mystery shopping scores and sales

Notes: This figure plots point estimates and 90% confidence intervals for the coefficient on log sales in the various regression specifications where the dependent variable is the standardized mystery shopping score (coefficients plotted in black) or its standardized components: cleanliness (dark gray), product (medium gray), and personnel (light gray). Circles refer to coefficients from the regressions of mystery shopping scores on log sales with controls for the time of the visit (morning until 11:59am, midday until 2:59pm, evening thereafter), squares to the regressions with time and month fixed effects, diamonds to the regressions with time, month and mystery shopper fixed effects, and triangles to the regressions with time, month, shopper and site fixed effects. As we do not have data on mystery shoppers for firm 3, the estimates coded with an "X" refers to the regression with time, month and site fixed effects. Standard errors are clustered at the site level. More detailed regression output is available in Tables 9, 10, and 11 in Appendix A.

Figure 2: Bonus payments to site supervisors

Notes: This histogram plots the size of bonuses paid to site supervisors in percent of their wages. For Firm 2, we express monthly bonuses in percent of monthly wages. For Firm 3, we express quarterly bonuses in percent of quarterly wages. If there is more than one site supervisor in a site-month or site-quarter, we sum up the bonus payments and wage payments of all site supervisors.

1. Using different lags and moving averages of MS scores (up to 12 months), we find no correlation for Firms 1 and 2, and a small negative correlation for Firm 3 (detailed results in Tables 13, 14 and 15 in Appendix A), none of which results supports this alternative explanation.

4.3 Mystery shopping scores and site incentives (hypothesis H3)

The franchise Firm 1 has clear rules governing its incentives that are stipulated in the franchise contract. In addition to the residual claim on 93 to 95% of net sales, the following MS-based incentives apply. Should a site's MS score fall below 90%, a conversation with the firm's representative focusing on service delivery will follow. Each time a site is scored below 75%, a fine of 50 Euros is paid. A further fine of 50 Euros is charged when the "eat in or take away?" question is not asked. If a site is scored below 75% more than once in a row, the franchisee has to submit a binding plan for improving service quality. Failure to agree on the plan or to follow up on it results in a cancellation of the franchise contract, which is a real danger, given a large initial and firm-specific investment (according to the firm's website, at least 30k Euros). Firm 1's setting does not entirely fit that of hypothesis H3 (there is no discretion in administering incentives). Still, by fining sub-standard performance rather than rewarding above-standard performance, Firm 1's policies are aimed at ensuring service uniformity rather than service excellence, and are thus consistent with our theory.

Turning to the integrated Firms 2 and 3, their site supervisor incentives are, predictably, weaker and more discretionary rather than rules-based. They are weaker because site supervisors are salaried employees rather than residual claimants, and are discretionary because the integrated firms have controls and information about the sites that they can use in addition to hard performance data. Figure 2 plots the distributions of the site supervisor bonus in Firms 2 (paid monthly) and 3 (paid quarterly) as a percentage of salary. Non-zero bonuses are awarded 65% of the time in Firm 2 and 44% of the time in Firm 3. Conditional on the receipt, the bonus size varies from less than 10% to over 20% of the salary, averaging at a substantial 13.4% in Firm 2 and 11% in Firm 3.

We now turn to examining the factors affecting the receipt and the size of a bonus as informed by the incentive policies practiced in the firms (recall Section 3), as well as by our model, namely: sales and personnel costs, the respective targets, and MS scores, plus additional control for the baseline wage (Firm 2 seems to use it as the basis for the bonus, see Figure 2) and time fixed effects (firms may have different bonus budgets in different time periods). We begin by running two independent regressions: one, probit, for the receipt of the bonus, the other, linear, for log size of the bonus (columns 1 and 4 in Table 4). We next combine them into the two-stage Heckman sample selection model which we estimate allowing for the error terms from the two equations to be correlated. We first estimate the model without exclusion restrictions (columns 2 and 5 in Table 4), and then impose exclusion restrictions to improve identifiablity (columns 3 and 6 in Table 4).[17]

The regression results presented in Table 4 show that the likelihood of receiving a bonus increases with sales above the target and with MS scores in both firms. Conditional on the receipt, the bonus size in Firm 2 increases with sales above the target and decreases with personnel costs above the target. Baseline wage level also matters, reflecting Firm 2's practice to tie bonus size to the baseline salary level. MS scores are unimportant for bonus size in Firm 2. In Firm 3, the likelihood of receiving a bonus grows with sales above the target, baseline salary, and with the MS score, but the link with personnel costs is not robust to specification. Bonus size increases with sales, both absolute and relative to the target, and is also positively affected by the MS score. The

[17]Heckman selection models have been widely used in strategy research to address non-random selection of observations into a study sample and rectify the biases in regression estimates caused by this selection. We follow the estimation method outlined in Certo et al. (2016) and Wolfolds and Siegel (2019), who also survey its applications.

Table 4: Determinants of bonus payments to site supervisors

	Firm 2			Firm 3		
	Simple	Heckman		Simple	Heckman	
	(1)	(2)	(3)	(4)	(5)	(6)
Linear regression	*Ln(bonus in €)*					
Lagged mystery shopping score	0.009***	0.009		0.122***	0.054***	0.047***
	(0.003)	(0.008)		(0.010)	(0.016)	(0.016)
Lagged ln(sales)	-0.135	-0.135		0.765***	0.722***	0.479***
	(0.092)	(0.092)		(0.146)	(0.178)	(0.028)
Lagged ln(sales)	2.565***	2.563***	2.422***	0.651***	0.545***	0.463***
-ln(sales target)	(0.229)	(0.231)	(0.209)	(0.236)	(0.177)	(0.180)
Lagged ln(personnel costs)	0.191**	0.190*	0.050	-0.328**	-0.203	
	(0.096)	(0.097)	(0.040)	(0.157)	(0.194)	
Lagged ln(personnel costs)	-0.835***	-0.834***	-0.707***	0.075	-0.255	
-ln(personnel costs target)	(0.096)	(0.096)	(0.078)	(0.160)	(0.193)	
Ln(supervisor wage)	0.470***	0.471***	0.449***	0.186	-0.036	
	(0.116)	(0.115)	(0.111)	(0.153)	(0.161)	
Probit	*Received bonus (=1 if yes)*					
Lagged mystery shopping score	0.219***	0.219***	0.217***	0.206***	0.183***	0.182***
	(0.013)	(0.013)	(0.012)	(0.021)	(0.022)	(0.022)
Lagged ln(sales)	-0.101	-0.100		-0.006	-0.083	
	(0.321)	(0.321)		(0.311)	(0.283)	
Lagged ln(sales)	0.749*	0.748*	0.650**	0.910**	0.996***	1.165***
-ln(sales target)	(0.403)	(0.402)	(0.265)	(0.365)	(0.366)	(0.356)
Lagged ln(personnel costs)	0.514	0.514		-0.232	-0.090	
	(0.363)	(0.363)		(0.335)	(0.311)	
Lagged ln(personnel costs)	-0.357	-0.358		1.090***	0.897***	0.322
-ln(personnel costs target)	(0.356)	(0.356)		(0.361)	(0.347)	(0.237)
Ln(supervisor wage)	-0.122	-0.120		0.704***	0.706***	0.472***
	(0.306)	(0.306)		(0.237)	(0.226)	(0.180)
Clustered at	Site	Site	Site	Site	Site	Site
Number of clusters	169	169	169	241	241	241
Observations	2,523	2,523	2,523	1,805	1,805	1,805
Complete obs. in linear reg.	1,950	1,950	1,950	975	975	975
Complete obs. in probit	2,523	2,523	2,523	1,805	1,805	1,805

Notes: This table reports regression results for the receipt and level of bonuses awarded to site supervisors in Firms 2 and 3. The receipt of the bonus is modelled as a probit regression, and the log size of the bonus is modelled as a linear regression of the covariates informed by the firms' incentive policies (Section 3) and our model. Columns 1 and 4 report the results for the receipt and size of the bonus equations estimated independently. Columns 2 and 5 report the results of the two equations combined in a Heckman selection model without exclusion restrictions, and columns 3 and 6 report the same with exclusion restrictions informed by the results from the unrestricted specification. Lagged covariates are used to account for the delay in processing performance results: 2 months for Firm 2, 1 quarter for Firm 3. Using other lags results in considerably lower model fit (details available on request). Quarterly averages are used in the regressions for Firm 3. Firm 2's bonus data are available for the period Jan. 2012 to Oct. 2013; Firm 3's data are available for the entire period of observations. We include month (quarter) fixed effects in all equations in Firm 2 (Firm 3). Coefficient standard errors are clustered at the site level and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

results are consistent across different specifications for Firm 2 but differ for Firm 3 in that the effect of MS scores on bonus size decreases once we control for endogenous selection into the positive bonus subsample in the Heckman model regressions (columns 5 and 6).[18]

Table 5: The average marginal effects of performance indicators on the likelihood and size of site supervisors' incentive bonuses

	Increase in P(bonus)		Increase in bonus size	
	Standard deviations	Percentage points	Standard deviations	Percent
Within-site standard deviation increase	(1)	(2)	(3)	(4)
Panel A: Firm 2				
Mystery shopping score	0.403	19.173	-	-
	(0.014)	(0.671)	-	-
Sales	0.022	1.041	0.475	25.700
	(0.009)	(0.435)	(0.041)	(2.101)
Personnel costs	-	-	-0.205	-11.473
	-	-	(0.028)	(1.445)
Panel B: Firm 3				
Mystery shopping score	0.395	19.726	0.258	17.068
	(0.044)	(2.180)	(0.089)	(5.908)
Sales	0.098	4.913	0.202	14.286
	(0.028)	(1.391)	(0.041)	(2.692)
Personnel costs	-	-	-	-
	-	-	-	-

Notes: The average marginal effects are calculated based on the regression estimates in columns 3 and 6 of table 4. In a regression model $y = f(x, \beta)$ estimated on a sample of n observations, the average marginal effect of a change Δ (1sd in our application) in a regressor x is the average of the implied changes in the outcome across the observations, or $\bar{m} = \frac{1}{n}\sum_{i=1}^{n}\left[f(x_i + \Delta, \beta) - f(x_i, \beta)\right]$, whose standard error is calculated using the Delta method. Notice that $\bar{m} = \Delta\beta$ in a linear regression model, and is thus constant and immediately available from regression output. This is not the case for nonlinear models such as the Heckman selection model that we estimate. The marginal effects are calculated with respect to the probability of receiving a bonus (columns 1 and 2), and bonus size (columns 3 and 4). If there is no entry in a cell, this is because the corresponding variable is not significant in the corresponding equation of the Heckman model in columns 3 and 6 in Table 4. Clustered standard errors are in parentheses.

To better relate the regression results in Table 4 to our hypothesis H3 (MS affects the likelihood of a bonus whose size increases with sales), we calculate the implied average marginal effects of a one-standard-deviation (1sd) increase in log sales and MS scores within sites on the likelihood and size of the bonus. Table 5 reports the results and

[18]The consistency of the results across specifications allowing and not allowing for non-random selection into the positive bonus subsample for Firm 2 may look surprising given that receiving a bonus is clearly endogenous to performance results. It is worth noting, however, that endogenous sample selection is the necessary but not sufficient condition for the sample selection bias (Certo et al., 2016, pp. 2647-49). Our interpretation of the consistency of the results across the specifications for Firm 2 is that, conditional on the observed performance results, the other, unobservable, factors affecting the receipt and the size of the bonus are uncorrelated with each other.

provides technical detail of the estimation procedure. For Firm 2, a 1sd increase in the MS score increases the likelihood of receiving the bonus by 19 percentage points (ppts), or 0.4sd. The effect of a 1sd increase in sales is much smaller: 1 ppts, or 0.02sd. The effects and their relative size are not very different for Firm 3: 20 ppts (0.4 sd) and 5% (0.1 sd) higher likelihood of a bonus with 1sd increase in MS score and sales, respectively. Thus, MS score has a major effect on the receipt of a bonus in both firms, whereas sales, though statistically significant, are much less important.

The relative importance of MS score and sales for bonus size is different. A 1sd increase in MS score in Firm 2 produces a statistically and economically insignificant increase in bonus size. The same increase in log sales, on the other hand, results in a 26% higher bonus. In Firm 3, however, MS score continues to significantly affect bonus size, with a 1sd increase in it resulting in a 17% higher bonus, which is comparable with the effect of a 1sd increase in sales: 14%.

Finally, in Table 16 in the appendix A, we document that the results reported in Table 4 are robust to accounting for missing MS observations in Firm 2.[19] In addition, we show in Table 17 in the appendix A that the results from Table 4 do not depend on the way we aggregate monthly MS scores in the quarterly bonus regressions for Firm 3.[20]

Overall, the results presented in this section are in line with our hypothesis H3: MS scores affect the likelihood of a bonus whose size is driven by sales. In Firm 1, site supervisors retain most of the net sales but get fined for below-standard MS scores. In Firms 2 and 3, MS scores are the main determinant of the receipt of a bonus whose size depends on sales. The results for Firm 3 are different in that MS score exerts a major influence on size as well as the likelihood of a bonus. While the effect of MS scores on the bonus size does not contradict H3, it cannot be rationalized within our model which predicts that incentives should encourage service uniformity. Still, despite incentives monotonically increasing with MS scores in Firm 3, they do not seem to cause heterogeneous MS results: indeed, it is Firm 3 where MS score has the lowest variance (Table 3).

[19] The lack of correlation between the likelihood of a missing MS visit and site characteristics (Section 4.2, Table 12 in Appendix A) allows us to estimate a relatively simple model where we impute missing MS scores with the sample mean (column 1), a linear prediction (column 2) or the predicted value obtained from an elastic net regression of the MS score on a number of site variables and their interactions (column 3), and add a dummy for the imputed observations as an additional control in all specifications. The estimates are very similar across the different specifications, suggesting the robustness of our main results to missing MS observations.This said, a missing MS score is not a mere statistical nuisance. Even though a missing MS score does not affect the relationships of our main theoretical interest or the size of the bonus received, the average marginal effect estimates based on the results in Table 16 in the appendix A (notes to Table 5 give technical detail) suggest that sites with a missing MS score are 6 to 7 percentage points less likely to receive a bonus than observationally equivalent sites with a non-missing MS score. This difference is another reflection of the importance of MS scores for site incentives.

[20] Aggregation may matter, for instance, when decision makers attach special importance to unusually high or low MS scores in the quarter preceding the bonus decision. In our case, all results remain when we use quarter-minimum MS scores instead of quarter-averages (Appendix Table 17).

5 Discussion and conclusion

In this study, we have linked the popular practice of MS to the agency problems that multi-site firms face in implementing their strategies. To better understand and operationalize this link, we have developed a theoretical framework that derives the optimal strategy for a multi-site firm in the presence of cross-site reputation spillovers: spend a fixed amount of resources to maintain service standards and the rest on stimulating sales. We have shown that the agency problems emerge because this strategy is not optimal from the perspective of individual sites. Consequently, a mechanism that ensures the firm's strategy is implemented by the sites is required. We have presented MS as part of such mechanism that generates service performance metrics used by the firm to incentivize its sites to adhere to its optimal strategy.

Our framework generates three testable hypotheses: lower variation in MS scores than in sales (H1), low correlation between MS scores and sales (H2), the likelihood of an incentive bonus depending on MS scores, and bonus size depending on sales (H3). All these hypotheses receive robust empirical support. In all of our three study firms, MS scores vary much less across sites than do sales, and there is no economically important correlation between MS scores and sales. The likelihood of a site receiving an incentive bonus is determined by its MS score, and the size of the bonus depends on sales (and MS score in case of Firm 3).

Our study relates to several strands of the literature and has important practical implications. Our main contribution is to strategy research, and lies in conceptualizing MS as a strategic management practice in multi-site firms. Why is it important to bring MS closer to the strategy scholars' attention? We show in the literature review (Section 2.1) that several aspects of MS as a practice fit Nag, Hambrick and Chen (2007)'s definition of "strategic"; most importantly, its use by multi-site firms in dealing with the agency problems that we describe in detail in Sections 2.2 and 2.2.2. Agency problems in multi-site firms have been covered in the strategy literature that has so far focused on franchise firms (e.g., Sorenson and Sørensen 2001; Lafontaine and Shaw 2005; Vroom and Gimeno 2007; Barthélemy 2008; Ater and Rigbi 2015). Yet, we have not been able to find a single strategy paper that would study MS in this or other contexts.

This is an important research gap because it limits our understanding of the practices firms use to address agency issues, leaving some empirical puzzles unresolved. For example, the low to zero correlations between MS and other performance indicators found in the earlier studies were attributed to the suboptimal reliability of MS scores (Finn and Kayandé 1999; Blessing and Natter 2019). However, if MS scores are unreliable, why do firms use them not only for performance evaluation but also for incentives purposes?

Our explanation, informed by our theory and supported by empirical results (Section 4.3), is that MS scores and sales are uncorrelated because firms strategically choose to implement uniform service standards, which they enforce with MS-based incentives, while accepting variation in sales and rewarding it with progressive bonuses.

Beyond strategy research, our study also speaks to a literature in management accounting examining the uses and effects of action controls (Widener, Shackell and Demers 2008; Campbell, Epstein and Martinez-Jerez 2011; Merchant and Van der Stede 2017; Arnold and Posch 2020). This literature briefly mentions that MS serves as an action control, and finds that actions controls, such as accountability through exception reports, increase adherence to standards, and reduce heterogeneity in effort allocation (Campbell, Epstein and Martinez-Jerez 2011). Our findings are consistent with this view, and suggest, additionally, that explicit incentives may be one way to integrate action controls with other control mechanisms and to enable action controls to be effective. Relatedly, our results are in line with the literature on the use of multiple controls that argues that multiple controls should be used by firms and that combining outcome and action controls is preferred to solely relying on either type (Kreutzer, Walter and Cardinal 2015; Sihag and Rijsdijk 2019).

More broadly, our work relates to the research on the nexus between strategy and accounting that emphasizes the importance of aligning HR management and performance measurement practices with the technological and strategic environment of the firm (Kaplan and Norton 1996; Baron and Kreps 1999; Van der Stede, Chow and Lin 2006; Gans and Ryall 2017; Abernethy, Kuang and Qin 2019). In terms of this research, our results show that MS-based incentives are used on top of sales-based incentives to align sites' interests with the firm's strategy balancing local responsiveness with chain uniformity. Having multiple performance goals alone is not enough for achieving strategic alignment, however. Studies on organizations with multiple performance goals (Barthélemy 2008; Ethiraj and Levinthal 2009; Obloj and Sengul 2020) warn of performance losses due to complex trade-offs involved in trying to reach multiple goals simultaneously. Setting a uniform service standard and incentivizing it accordingly helps ease the tension between the conflicting demands on employee resources and reduce multitasking (too much focus on more strongly incentivized activity; Holmström and Milgrom 1991; Feltham and Xie 1994).

Turning to practical implications, our work shows that sales-based incentives alone are not enough to support the optimal allocation of site resources between sales- and service-focused activities from the multi-site firm's perspective. Additional incentives, based on service performance, are necessary. Firms are advised to practice MS and use its results for evaluation and reward purposes together with other performance indicators. A lack of correlation between MS scores and other performance indicators

that may seem related (e.g., sales) should not be taken to imply that MS is an unreliable performance measurement tool. It may instead be evidence that MS as a strategic management practice is effective in encouraging sites to adhere to service standards desired by the firm.

Our study is not without limitations. The theoretical framework lacks a formal characterization of the optimal incentive scheme the multi-site firm in our model should use to ensure the implementation of its strategy. The scheme we propose – MS affecting the likelihood of a bonus, sales affecting bonus size – does imply uniform service performance and varying sales, which we observe, but it is not necessarily the optimal one. Another limitation of our study lies in not being able to observe the causal effect of MS on performance results. Without this information, we have had to rely on other studies (e.g., Banerjee et al. 2021) or anecdotes and quasi-experiments in our own data, like the change in MS scoring system described in footnote 15, in arguing for the effectiveness of MS in influencing sites' actions. Relatedly, there is a question of external validity of our findings, given that our sample consists of only three study firms in similar lines of business (food retail). However, the prominence of the food retail sector in the modern economy and similarity of our main results across the firms makes us believe our findings are applicable beyond our study sample.

Notwithstanding these limitations, we believe our work is noteworthy and is a useful start of a potentially fruitful line of research into strategic aspects of MS as well as other performance measurement practices in multi-site firms where reputation spillovers are important.

References

Abernethy, Margaret A., Yu Flora Kuang, and Bo Qin. 2019. "The relation between strategy, CEO selection, and firm performance." *Contemporary Accounting Research*, 36(3): 1575–1606.

Anderson, Eugene W., and Vikas Mittal. 2000. "Strengthening the satisfaction-profit chain." *Journal of Service Research*, 3(2): 107–120.

Arnold, Markus C., and Arthur Posch. 2020. "The Use and Effects of Accountability and Job Autonomy when Results Controls are Irrelevant: Substitutes or Complements?" Working paper.

Arnold, Markus C., and Ivo D. Tafkov. 2019. "Managerial discretion and task interdependence in teams." *Contemporary Accounting Research*, 36(4): 2467–2493.

Ater, Itai, and Oren Rigbi. 2015. "Price control and advertising in franchising chains." *Strategic Management Journal*, 36(1): 148–158.

Baiman, Stanley, and Madhav V. Rajan. 1995. "The informational advantages of discretionary bonus schemes." *The Accounting Review*, 70(4): 557–579.

Baker, George. 2002. "Distortion and Risk in Optimal Incentive Contracts." *The Journal of Human Resources*, 37(4): 728–751.

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh. 2021. "Improving police performance in Rajasthan, India: Experimental evidence on incentives, managerial autonomy, and training." *American Economic Journal: Economic Policy*, 13(1): 36–66.

Baron, James N., and David M. Kreps. 1999. *Strategic Human Resources: Frameworks for General Managers.* New York:John Wiley & Sons Inc.

Barthélemy, Jérôme. 2008. "Opportunism, knowledge, and the performance of franchise chains." *Strategic Management Journal*, 29(13): 1451–1463.

Beck, Jeff, and Li Miao. 2003. "Mystery shopping in lodging properties as a measurement of service quality." *Journal of Quality Assurance in Hospitality & Tourism*, 4(1-2): 1–21.

Blessing, Gerald, and Martin Natter. 2019. "Do mystery shoppers really predict customer satisfaction and sales performance?" *Journal of Retailing*, 95(3): 47–62.

Bol, Jasmijn C., Gary Hecht, and Steven D. Smith. 2015. "Managers' discretionary adjustments: The influence of uncontrollable events and compensation interdependence." *Contemporary Accounting Research*, 32(1): 139–159.

Bol, Jasmijn C., Timothy M. Keune, Ella M. Matsumura, and Jae Yong Shin. 2010. "Supervisor discretion in target setting: An empirical investigation." *The Accounting Review*, 85(6): 1861–1886.

Bouwens, Jan, and Peter Kroos. 2017. "The interplay between forward-looking measures and target setting." *Management Science*, 63(9): 2868–2884.

Bradach, Jeffrey L. 1995. "Chains within chains: The role of multi-unit franchisees." *Journal of Marketing Channels*, 4(1-2): 65–81.

Bradach, Jeffrey L. 1997. "Using the plural form in the management of restaurant chains." *Administrative Science Quarterly*, 42(2): 276–303.

Bradach, Jeffrey L. 1998. *Franchise Organizations.* Harvard Business School Press.

Brickley, James A., and Frederick H. Dark. 1987. "The choice of organizational form: the case of franchising." *Journal of Financial Economics*, 18(2): 401–420.

Cameron, Kim S. 1986. "Effectiveness as Paradox: Consensus and Conflict in Conceptions of Organizational Effectiveness." *Management Science*, 32(5): 539–553.

Campbell, Dennis. 2008. "Nonfinancial performance measures and promotion-based incentives." *Journal of Accounting Research*, 46(2): 297–332.

Campbell, Dennis, Marc J. Epstein, and F. Asis Martinez-Jerez. 2011. "The learning effects of monitoring." *The Accounting Review*, 86(6): 1909–1934.

Certo, S. Trevis, John R. Busenbark, Hyun-soo Woo, and Matthew Semadeni. 2016. "Sample selection bias and Heckman models in strategic management research." *Strategic Management Journal*, 37(13): 2639–2657.

Cheo, Roland, Ge Ge, Geir Godager, Rugang Liu, Jian Wang, and Qiqi Wang. 2020. "The effect of a mystery shopper scheme on prescribing behavior in primary care: Results from a field experiment." *Health Economics Review*, 10(1): 1–19.

Correia, Sergio. 2015. "Singletons, cluster-robust standard errors and fixed effects: A bad mix." Technical note.

Erstad, Margaret. 1998. "Mystery shopping programmes and human resource management." *International Journal of Contemporary Hospitality Management*, 10(1): 34–38.

Ethiraj, Sendil K., and Daniel Levinthal. 2009. "Hoping for A to Z while rewarding only A: Complex organizations and multiple goals." *Organization Science*, 20(1): 4–21.

Fama, Eugene F., and Michael C. Jensen. 1983. "Separation of ownership and control." *Journal of Law and Economics*, 26(2): 301–326.

Feltham, Gerald A., and Jim Xie. 1994. "Performance measure congruity and diversity in multi-task principal/agent relations." *The Accounting Review*, 69(3): 429–453.

Finn, Adam. 2001. "Mystery Shopper Benchmarking of Durable-Goods Chains and Stores." *Journal of Service Research*, 3(4): 310–320.

Finn, Adam. 2007. "Doing a Double Take: Accounting for Occasions in Service Performance Assessment." *Journal of Service Research*, 9(4): 372–387.

Finn, Adam, and Ujwal Kayandé. 1999. "Unmasking a phantom: a psychometric assessment of mystery shopping." *Journal of Retailing*, 75(2): 195–217.

Friebel, Guido, and Michael Raith. 2010. "Resource allocation and organizational form." *American Economic Journal: Microeconomics*, 2(2): 1–33.

Friebel, Guido, Matthias Heinz, and Nikolay Zubanov. 2022. "Middle Managers, Personnel Turnover, and Performance: A Long-Term Field Experiment in a Retail Chain." *Management Science*, 68(1): 211–229.

Friebel, Guido, Matthias Heinz, Ingo Weller, and Nikolay Zubanov. 2021. "Workplace Productivity and Management Practices (Research in Labor Economics, Vol. 49)." , ed. S.W. Polachek, K. Tatsiramos, G. Russo and G. van Houten, Chapter Downsizing Announcements, Job Security Perceptions, and Worksite Performance, 179–205. Emerald Publishing Limited.

Friebel, Guido, Matthias Heinz, Miriam Krueger, and Nikolay Zubanov. 2017. "Team incentives and performance: Evidence from a retail chain." *American Economic Review*, 107(8): 2168–2203.

Friebel, Guido, Matthias Heinz, Mitchell Hoffman, and Nikolay Zubanov. 2022. "What Do Employee Referral Programs (ERPs) Do? Measuring the Direct and Overall Effects of a Management Practice." *Journal of Political Economy*, forthcomming.

Gans, Joshua, and Michael D. Ryall. 2017. "Value capture theory: A strategic management review." *Strategic Management Journal*, 38(1): 17–41.

Gibbons, Robert, and John Roberts. 2013. *The handbook of organizational economics.* Princeton University Press Princeton, NJ. Chapter 20.

Gibbs, Michael, Kenneth A. Merchant, Wim A. Van der Stede, and Mark E. Vargus. 2004. "Determinants and effects of subjectivity in incentives." *The Accounting Review*, 79(2): 409–436.

Gillis, William E, James G Combs, and Xiaoli Yin. 2020. "Franchise management capabilities and franchisor performance under alternative franchise ownership strategies." *Journal of Business Venturing*, 35(1).

Helm, Sabrina, and Risto T. Salminen. 2010. "Basking in reflected glory: Using customer reference relationships to build reputation in industrial markets." *Industrial Marketing Management*, 39(5): 737–743.

Heskett, James L., Thomas O. Jones, Gary W. Loveman, W. Earl Sasser, and Leonard A. Schlesinger. 1994. "Putting the service-profit chain to work." *Harvard Business Review*, 72(2): 164–174.

Holmström, Bengt, and Paul Milgrom. 1991. "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design." *Journal of Law, Economics, & Organization*, 7(special issue): 24–52.

Ingram, Paul, and Joel A.C. Baum. 1997. "Chain Affiliation and the Failure of Manhattan Hotels." *Administrative Science Quarterly*, 42(1): 68–102.

Iranzo, Susana, Fabiano Schivardi, and Elisa Tosetti. 2008. "Skill Dispersion and Firm Productivity: An Analysis with Employer-Employee Matched Data." *Journal of Labor Economics*, 26(2): 247–285.

Jacob, Steve, Nathalie Schiffino, and Benjamin Biard. 2018. "The mystery shopper: a tool to measure public service delivery?" *International Review of Administrative Sciences*, 84(1): 164–184.

Jensen, Michael C., and William H. Meckling. 1976. "Theory of the firm: Managerial behavior, agency costs and ownership structure." *Journal of Financial Economics*, 3(4): 305–360.

Jin, Ginger Zhe, and Phillip Leslie. 2009. "Reputational incentives for restaurant hygiene." *American Economic Journal: Microeconomics*, 1(1): 237–267.

Kaplan, Robert S., and David P. Norton. 1996. "Using the balanced scorecard as a strategic management system." *Harvard Business Review*, 47(1): 75–85.

Kaufmann, Patrick J., and Sevgin Eroglu. 1999. "Standardization and adaptation in business format franchising." *Journal of Business Venturing*, 14(1): 69–85.

Khashabi, Pooyan, Matthias Heinz, Nikolay Zubanov, Tobias Kretschmer, and Guido Friebel. 2021. "Market Competition and the Effectiveness of Performance Pay." *Organization Science*, 32(2): 334–351.

Kidwell, Roland E., Arne Nygaard, and Ragnhild Silkoset. 2007. "Antecedents and effects of free riding in the franchisor–franchisee relationship." *Journal of Business Venturing*, 22(4): 522–544.

Kreutzer, Markus, Jorge Walter, and Laura B. Cardinal. 2015. "Organizational control as antidote to politics in the pursuit of strategic initiatives." *Strategic Management Journal*, 36(9): 1317–1337.

Lafontaine, Francine, and Kathryn L. Shaw. 2005. "Targeting managerial control: evidence from franchising." *The RAND Journal of Economics*, 36(1): 131–150.

Latham, Gary P., Robert C. Ford, and Danny Tzabbar. 2012. "Enhancing employee and organizational performance through coaching based on mystery shopper feedback: A quasi-experimental study." *Human Resource Management*, 51(2): 213–229.

Lovallo, Dan, and Olivier Sibony. 2018. "Broadening the frame: How behavioral strategy redefines strategic decisions." *Strategy Science*, 3(4): 658–667.

Makofske, Matthew Philip. 2020. "Mandatory disclosure, letter-grade systems, and corruption: The case of Los Angeles County restaurant inspections." *Journal of Economic Behavior and Organization*, 172: 292–313.

Manthei, Kathrin, Dirk Sliwka, and Timo Vogelsang. 2021. "Performance pay and prior learning—evidence from a retail chain." *Management Science*, 67(11): 6998–7022.

Maritan, Catherine A., and Gwendolyn K. Lee. 2017. "Resource allocation and strategy." *Journal of Management*, 43(8): 2411–2420.

Meiseberg, Brinja. 2013. "The prevalence and performance impact of synergies in the plural form." *Managerial and Decision Economics*, 34(3-5): 140–160.

Merchant, Kenneth A., and Wim A. Van der Stede. 2017. *Management Control Systems - Performance Measurement, Evaluation and Incentives.* Harlow:Pearson Education.

Michael, Steven C. 2000. "The effect of organizational form on quality: the case of franchising." *Journal of Economic Behavior & Organization*, 43(3): 295–318.

Milgrom, Paul, and John Roberts. 1990. "The economics of modern manufacturing: Technology, strategy, and organization." *American Economic Review*, 80: 511–528.

MSPA. 2018. "Mystery Shopping – how big is the market." Available at: https://www.mspa-ea.org/news/newsitem/58-mystery-shopping-how-big-is-the-market.htm. Accessed July 1st, 2021.

Nag, Rajiv, Donald C. Hambrick, and Ming-Jer Chen. 2007. "What is strategic management, really? Inductive derivation of a consensus definition of the field." *Strategic Management Journal*, 28(9): 935–955.

Nyberg, Anthony J., Mark A. Maltarich, Dhuha "Dee" Abdulsalam, Spenser M. Essman, and Ormonde Cragun. 2018. "Collective pay for performance: A cross-disciplinary review and meta-analysis." *Journal of Management*, 44(6): 2433–2472.

Obloj, Tomasz, and Metin Sengul. 2020. "What do multiple objectives really mean for performance? Empirical evidence from the French manufacturing sector." *Strategic Management Journal*, 41(13): 2518–2547.

Rajan, Madhav V., and Stefan Reichelstein. 2006. "Subjective performance indicators and discretionary bonus pools." *Journal of Accounting Research*, 44(3): 585–618.

Shane, Scott. 1996. "Hybrid organizational arrangements and their implications for firm growth and survival: A study of new franchisors." *Academy of Management Journal*, 39(1): 216–234.

Sihag, Vikrant, and Serge A. Rijsdijk. 2019. "Organizational Controls and Performance Outcomes: A Meta-Analytic Assessment and Extension." *Journal of Management Studies*, 56(1): 91–133.

Sorenson, Olav, and Jesper B. Sørensen. 2001. "Finding the right mix: Franchising, organizational learning, and chain performance." *Strategic Management Journal*, 22(6-7): 713–724.

Torpey, Elka. 2016. "Mystery shopper." U.S. Bureau of Labor Statistics Career Outlook. Available at: https://www.bls.gov/careeroutlook/2016/youre-a-what/mystery-shopper.htm. Accessed March 12th, 2022.

Van der Stede, Wim A., Chee W. Chow, and Thomas W. Lin. 2006. "Strategy, choice of performance measures, and performance." *Behavioral Research in Accounting*, 18(1): 185–205.

Vroom, Govert, and Javier Gimeno. 2007. "Ownership form, managerial incentives, and the intensity of rivalry." *Academy of Management Journal*, 50(4): 901–922.

Wall Street Journal. 2001. "McDonald's Asks Mystery Shoppers What Ails Sales." Available at: http://www.sddt.com/News/article.cfm?SourceCode=20011217fr&_t=McDonalds+Asks+Mystery+Shoppers+What+Ails+Sales#.X70n-rMxmUk. Accessed November 24th, 2020.

Widener, Sally K., Margaret B. Shackell, and Elizabeth A. Demers. 2008. "The juxtaposition of social surveillance controls with traditional organizational design components." *Contemporary Accounting Research*, 25(2): 605–638.

Wilson, Alan M. 1998*a*. "The role of mystery shopping in the measurement of service performance." *Managing Service Quality: An International Journal*, 8(6): 414–442.

Wilson, Alan M. 1998*b*. "The use of mystery shopping in the measurement of service delivery." *Service Industries Journal*, 18(3): 148–163.

Wilson, Alan M. 2001. "Mystery shopping: Using deception to measure service performance." *Psychology & Marketing*, 18(7): 721–734.

Wolfolds, Sarah E., and Jordan Siegel. 2019. "Misaccounting for endogeneity: The peril of relying on the Heckman two-step method without a valid instrument." *Strategic Management Journal*, 40(3): 432–462.

Yin, Xiaoli, and Edward J. Zajac. 2004. "The strategy/governance structure fit relationship: Theory and evidence in franchising arrangements." *Strategic Management Journal*, 25(4): 365–383.

A Appendix - Additional results

Table 6: Firm 1 - Analysis of variance: Aggregate mystery shopping scores

	Partial SS	% of total SS	df	MS	F	Prob >F
Model	213,525	53.36%	723	295.33	4.96	0.0000
Site	45,180	11.29%	353	127.99	2.15	0.0000
Visit month	30,409	7.60%	12	2,534.10	42.55	0.0000
Mystery shopper	70,318	17.57%	356	197.52	3.32	0.0000
Visit time	4,548	1.14%	2	2,273.75	38.18	0.0000
Residual	186,602	46.64%	3,113	59.56		
Total	40,0128	100.00%	3,856	103.77		
Observations	3,857					
R-squared	0.534					
Adjusted R-squared	0.426					

Notes: This table reports results from an analysis of variance of the aggregate mystery shopping score. The variable visit time takes the value 1 if the visit took place in the morning (until 11 A.M.), takes the value 2 if the visit took place during midday (from 11 A.M. to 2.59 P.M.), and takes the value 3 if the visit took place in the evening (from 3 P.M.). We arrive at the sample size as follows: From the total 8,115 mystery shopping scores (table 2), we know the identity of mystery shoppers for 4,939 mystery shopping scores. Moreover, including the time of the mystery shopping visit reduces the sample size to 3,857.

Table 7: Firm 2 - Analysis of variance: Aggregate mystery shopping scores

	Partial SS	% of total SS	df	MS	F	Prob >F
Model	40,339	35.67%	608	66.35	3.10	0.0000
Site	8,686	7.68%	230	37.77	1.76	0.0000
Visit month	1,714	1.52 %	23	74.54	3.48	0.0000
Mystery shopper	24,007	21.23%	353	68.01	3.17	0.0000
Visit time	1,202	1.06%	2	601.19	28.05	0.0000
Residual	72,745	64.33%	3,394	21.43		
Total	113,084	100.00%	4,002	28.26		
Observations	4,003					
R-squared	0.357					
Adjusted R-squared	0.242					

Notes: This table reports results from an analysis of variance of the aggregate mystery shopping score. The variable visit time takes the value 1 if the visit took place in the morning (until 11 A.M.), takes the value 2 if the visit took place during midday (from 11 A.M. to 2.59 P.M.), and takes the value 3 if the visit took place in the evening (from 3 P.M.). We arrive at the sample size as follows: From the total 4,965 mystery shopping scores (table 2), we do not know the identity of mystery shoppers for 962 mystery shopping scores.

Table 8: Firm 3 - Analysis of variance: Aggregate mystery shopping scores

	Partial SS	% of total SS	df	MS	F	Prob >F
Model	16,146	17.12%	276	58.50	5.57	0.0000
Site	8,413	8.92%	240	35.05	3.34	0.0000
Visit month	7,243	7.68%	34	213.03	20.29	0.0000
Visit time	395	0.42%	2	197.59	18.82	0.0000
Residual	78,184	82.88%	7,445	10.50		
Total	94,330	100.00%	7,721	12.22		
Observations	7,722					
R-squared	0.171					
Adjusted R-squared	0.140					

Notes: This table reports results from an analysis of variance of the aggregate mystery shopping score. The variable visit time takes the value 1 if the visit took place in the morning (until 11 A.M.), takes the value 2 if the visit took place during midday (from 11 A.M. to 2.59 P.M.), and takes the value 3 if the visit took place in the evening (from 3 P.M.). We arrive at the sample size as follows: From the total 8,654 mystery shopping scores (table 2), we do not know the time of the mystery shopping visit for 932 mystery shopping scores.

Table 9: Firm 1 - Regressing mystery shopping scores on daily sales

	Standardized mystery shopping score				*Standardized cleanliness score*				*Standardized product score*				*Standardized personnel score*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Ln(daily sales)	0.086**	0.092**	0.081**	0.060	-0.081**	-0.075**	-0.111***	-0.178**	0.252***	0.259***	0.260***	0.265***	0.049	0.051	0.065	0.091
	(0.041)	(0.041)	(0.038)	(0.078)	(0.037)	(0.037)	(0.031)	(0.070)	(0.040)	(0.040)	(0.038)	(0.084)	(0.045)	(0.045)	(0.044)	(0.080)
Morning	0.362***	0.375***	0.340***	0.355***	0.171***	0.203***	0.172***	0.199***	0.573***	0.580***	0.554***	0.582***	0.094*	0.086*	0.065	0.044
	(0.053)	(0.052)	(0.049)	(0.049)	(0.049)	(0.051)	(0.050)	(0.052)	(0.049)	(0.049)	(0.049)	(0.051)	(0.052)	(0.051)	(0.051)	(0.051)
Midday	0.168***	0.217***	0.222***	0.241***	0.006	0.067*	0.061*	0.080**	0.405***	0.431***	0.434***	0.451***	0.010	0.030	0.036	0.041
	(0.041)	(0.039)	(0.036)	(0.035)	(0.039)	(0.037)	(0.035)	(0.035)	(0.041)	(0.041)	(0.041)	(0.042)	(0.040)	(0.038)	(0.038)	(0.038)
Month fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Shopper fixed effects	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Site fixed effects	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes
Clustered at	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site
Number of clusters	346	346	346	346	346	346	346	346	346	346	346	346	346	346	346	346
Observations	3,643	3,643	3,643	3,643	3,643	3,643	3,643	3,643	3,643	3,643	3,643	3,643	3,643	3,643	3,643	3,643
Adjusted R-squared	0.017	0.166	0.363	0.422	0.006	0.197	0.368	0.410	0.064	0.104	0.271	0.310	0.001	0.102	0.221	0.295

Notes: This table reports results of regressions of mystery shopping performance on sales. The data are at the mystery shopping visit level. The dependent variables are the z-scored aggregate mystery shopping score, and the three z-scored mystery shopping sub scores (cleanliness, product, and personnel). The main independent variable is the natural logarithm of the daily sales performance in Euros. We include as control variables two indicators for the time of the mystery shopping visit. *Morning* takes the value one if the mystery shopping visit took place until 11.59 A.M. *Midday* takes the value one if the mystery shopping visit took place between 12 P.M. and 2.59 P.M. For each mystery shopping score, we first present a simple OLS regression, and then we subsequently add month fixed effects, mystery shopper fixed effects, and site fixed effects. Coefficient standard errors are clustered at the site level and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01. We arrive at the final sample size as follows: From the total 8,115 mystery shopping scores (table 2), we can match 6,079 mystery shopping scores with daily sales data. Controlling for mystery shopper fixed effects reduces the sample to 3,725 mystery shopping scores. Finally, we drop 82 singleton observations to avoid over-stating statistical significance (Correia 2015).

Table 10: Firm 2 - Regressing mystery shopping scores on monthly sales

	Standardized mystery shopping score				*Standardized cleanliness score*				*Standardized product score*				*Standardized personnel score*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Ln(monthly sales)	0.017	0.015	0.040	-0.132	-0.179***	-0.178***	-0.181***	0.147	0.146***	0.138***	0.150***	-0.117	0.002	0.004	0.027	-0.148
	(0.053)	(0.053)	(0.051)	(0.224)	(0.038)	(0.038)	(0.041)	(0.199)	(0.040)	(0.040)	(0.035)	(0.227)	(0.054)	(0.054)	(0.053)	(0.236)
Morning	0.288***	0.278***	0.249***	0.284***	0.167***	0.181***	0.147***	0.143***	0.283***	0.265***	0.245***	0.259***	0.192***	0.187***	0.161***	0.198***
	(0.048)	(0.048)	(0.050)	(0.050)	(0.038)	(0.039)	(0.043)	(0.043)	(0.041)	(0.041)	(0.042)	(0.045)	(0.049)	(0.050)	(0.053)	(0.052)
Midday	0.263***	0.271***	0.218***	0.244***	0.076*	0.094**	0.062	0.063	0.290***	0.288***	0.246***	0.252***	0.178***	0.187***	0.144***	0.175***
	(0.039)	(0.038)	(0.038)	(0.038)	(0.040)	(0.040)	(0.041)	(0.040)	(0.037)	(0.036)	(0.035)	(0.035)	(0.039)	(0.039)	(0.039)	(0.039)
Month fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Shopper fixed effects	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Site fixed effects	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes
Clustered at	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site
Number of clusters	225	225	225	225	225	225	225	225	225	225	225	225	225	225	225	225
Observations	3,882	3,882	3,882	3,882	3,882	3,882	3,882	3,882	3,882	3,882	3,882	3,882	3,882	3,882	3,882	3,882
Adjusted R-squared	0.016	0.040	0.207	0.242	0.008	0.010	0.132	0.135	0.024	0.038	0.147	0.151	0.007	0.032	0.181	0.222

Notes: This table reports results of regressions of mystery shopping performance on sales. The data are at the mystery shopping visit level. The dependent variables are the z-scored aggregate mystery shopping score, and the three z-scored mystery shopping sub scores (cleanliness, product, and personnel). The main independent variable is the natural logarithm of the monthly sales performance in Euros. We include as control variables two indicators for the time of the mystery shopping visit. *Morning* takes the value one if the mystery shopping visit took place until 11.59 A.M. *Midday* takes the value one if the mystery shopping visit took place between 12 P.M. and 2.59 P.M. For each mystery shopping score, we first present a simple OLS regression, and then we subsequently add month fixed effects, mystery shopper fixed effects, and site fixed effects. Coefficient standard errors are clustered at the site level and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01. We arrive at the final sample size as follows: From the total 4,965 mystery shopping scores (table 2), we can match 4,932 mystery shopping scores with monthly sales data. Controlling for mystery shopper fixed effects reduces the sample to 3,974 mystery shopping scores. Finally, we drop 92 singleton observations to avoid over-stating statistical significance (Correia 2015).

Table 11: Firm 3 - Regressing mystery shopping scores on monthly sales

	Standardized mystery shopping score		
	(1)	(2)	(3)
Ln(monthly sales)	-0.071***	-0.076***	-0.261***
	(0.023)	(0.023)	(0.088)
Morning	-0.144***	-0.164***	-0.145***
	(0.027)	(0.027)	(0.026)
Midday	-0.043	-0.039	-0.032
	(0.028)	(0.027)	(0.028)
Month fixed effects	No	Yes	Yes
Shopper fixed effects	No	No	No
Site fixed effects	No	No	Yes
Clustered at	Site	Site	Site
Number of clusters	241	241	241
Observations	7,223	7,223	7,223
Adjusted R-squared	0.007	0.085	0.145

Notes: This table reports results of regressions of mystery shopping performance on sales. The data are at the mystery shopping visit level. The dependent variables is the z-scored aggregate mystery shopping score. The main independent variable is the natural logarithm of the monthly sales in Euros. We include as control variables two indicators for the time of the mystery shopping visit. *Morning* takes the value one if the mystery shopping visit took place until 11.59 A.M. *Midday* takes the value one if the mystery shopping visit took place between 12 P.M. and 2.59 P.M. For each mystery shopping score, we first present a simple OLS regression, and then we subsequently add month fixed effects, mystery shopper fixed effects, and site fixed effects. Coefficient standard errors are clustered at the site level and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01. We arrive at the final sample size as follows: From the total 8,654 mystery shopping scores (table 2), we can match 8,155 mystery shopping scores with monthly sales data. Controlling for the time of the mystery shopping visit reduces the sample to 7,223 mystery shopping scores.

Table 12: Firm 2 - Correlation between mystery shopping scores and sales (Heckman)

	(1)
Linear regression	*Standardized mystery shopping score*
Ln(sales)	-0.021
	(0.253)
Morning	0.377***
	(0.063)
Midday	0.250***
	(0.045)
Probit	*Received bonus (=1 if yes)*
Lagged standardized mystery shopping score	0.060**
	(0.025)
Lagged ln(sales) - ln(sales target)	0.341
	(0.298)
Lagged ln(personnel costs) - ln(personnel costs target)	0.152
	(0.183)
Store size	-0.028
	(0.023)
Located in big town	-0.086
	(0.066)
Renovated in previous month	-0.335
	(0.314)
Month fixed effects	Both equations
Site fixed effects	Behavioral equation
Shopper fixed effects	Behavioral equation
Clustered at	Site
Number of clusters	186
Observations	3,182
Complete observations in linear regression	2,589
Complete observations in probit	3,182

Notes: This table reports Heckman regression results where we document the correlation between mystery shopping scores and sales, taking into account a potential selection of sites into being visited by mystery shopper. In the probit equation, the dependent variable is an indicator variable whether a site was visited by a mystery shopper in a month. We include as independent variables the z-scored mystery shopping score, the difference between the natural logarithm of actual sales and the natural logarithm of target sales as well as the difference between the natural logarithm of actual personnel costs and the natural logarithm of target personnel costs. All variables are monthly variables and lagged by one month (e.g., because visit schedules may be based on last month's performance). We average the mystery shopping scores of a site when there is more than one mystery shopping visit to a site in a month. In firm 2, there are 29 (six) site-months with two (three) visits to a site. Furthermore, we include as independent variables the site size and two indicator variables, one capturing whether a site is located in a big town (>100,000 inhabitants) and one capturing if a site was renovated in the previous month. We calculate site size as the principal component of a principal component analysis where the input variables are a site's monthly sales, monthly customers, and monthly hours worked. In the linear regression equation, the dependent variable is the z-scored mystery shopping score. We include as independent variables the same variables as in our main specifications in table 10, i.e. the logarithm of monthly sales performance, and two indicators capturing the time of a mystery shopping visit. *Morning* takes the value one if the mystery shopping visit took place until 11.59 A.M. *Midday* takes the value one if the mystery shopping visit took place between 12 P.M. and 2.59 P.M. In the behavioral equation, we also include site fixed effects, and mystery shopper fixed effects. Month fixed effects are included in both equations. Coefficient standard errors are clustered at the site level and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table 13: Firm 1 - Regressing sales on lagged mystery shopping scores

	Ln(monthly sales)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Standardized mystery shopping score	-0.010	-0.000								
	(0.008)	(0.002)								
1-month lag of standardized mystery shopping score	-0.003	0.002								
	(0.007)	(0.002)								
2-month lag of standardized mystery shopping score	0.002	0.000								
	(0.007)	(0.002)								
3-month moving average of standardized mystery shopping score			-0.010	0.002						
			(0.021)	(0.004)						
6-month moving average of standardized mystery shopping score					-0.032	0.004				
					(0.035)	(0.007)				
9-month moving average of standardized mystery shopping score							-0.057	0.014		
							(0.047)	(0.010)		
12-month moving average of standardized mystery shopping score									-0.071	0.019
									(0.058)	(0.016)
Ln(monthly hours worked)	0.498***	0.051***	0.498***	0.051***	0.500***	0.042***	0.508***	0.036***	0.510***	0.032**
	(0.052)	(0.012)	(0.052)	(0.012)	(0.056)	(0.011)	(0.057)	(0.012)	(0.063)	(0.015)
Month fixed effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Site fixed effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Clustered at	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site
Number of clusters	294	294	294	294	283	283	268	268	224	224
Observations	4,504	4,504	4,504	4,504	3,556	3,556	2,681	2,681	1,881	1,881
Adjusted R-squared	0.431	0.968	0.431	0.968	0.427	0.972	0.430	0.977	0.413	0.978

Notes: This table reports results of a regression of sales on lagged mystery shopping performance. The data are at the site-month level. The dependent variable is the natural logarithm of monthly sales in Euros. In columns 1 and 2, we include as main independent variable the z-scored mystery shopping performance of the current month, the last month, and the second to last month. We average mystery shopping scores prior to z-scoring if there were multiple mystery shopping visits in a site-month. In column 3 and 4, we include as main independent variable the 3-month moving average of z-scored mystery shopping performance calculated as the average of all z-scored mystery shopping scores collected for a site in the current, the last, and second to last month. In columns 5 to 10, we include as main independent variable the 6-month, 9-month, or 12-month moving average of the z-scored mystery shopping performance. We calculate these variables analogously to the 3-month moving average (i.e. starting with the current month). In all specifications, we control for the natural logarithm of the monthly hours worked in a site. The odd columns refer to simple linear regressions. The even columns refer to regressions where we include site fixed effects and month fixed effects. Coefficient standard errors are clustered at the site level and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01. We include all site-months for which we have data on the dependent variable, and the independent variables. Furthermore, we drop singleton observations to avoid over-stating statistical significance (Correia 2015).

Table 14: Firm 2 - Regressing sales on lagged mystery shopping scores

	Ln(monthly sales)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Standardized mystery shopping score	0.002	-0.001								
	(0.006)	(0.002)								
1-month lag of standardized mystery shopping score	0.008	0.002								
	(0.006)	(0.002)								
2-month lag of standardized mystery shopping score	0.008	0.001								
	(0.006)	(0.001)								
3-month moving average of standardized mystery shopping score			0.018	0.002						
			(0.016)	(0.004)						
6-month moving average of standardized mystery shopping score					0.033	-0.000				
					(0.035)	(0.008)				
9-month moving average of standardized mystery shopping score							0.059	0.008		
							(0.065)	(0.018)		
12-month moving average of standardized mystery shopping score									0.172*	0.037
									(0.099)	(0.030)
Ln(monthly hours worked)	0.717***	0.002	0.717***	0.002	0.748***	-0.017	0.729***	-0.048	0.827***	-0.078
	(0.112)	(0.021)	(0.112)	(0.021)	(0.103)	(0.033)	(0.120)	(0.049)	(0.133)	(0.069)
Month fixed effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Site fixed effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Clustered at	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site
Number of clusters	193	193	193	193	165	165	118	118	57	57
Observations	2,932	2,932	2,932	2,932	1,546	1,546	778	778	344	344
Adjusted R-squared	0.481	0.965	0.481	0.965	0.487	0.966	0.464	0.979	0.507	0.987

Notes: This table reports results of a regression of sales on lagged mystery shopping performance. The data are at the site-month level. The dependent variable is the natural logarithm of monthly sales in Euros. In columns 1 and 2, we include as main independent variable the z-scored mystery shopping performance of the current month, the last month, and the second to last month. We average mystery shopping scores prior to z-scoring if there were multiple mystery shopping visits in a site-month. In column 3 and 4, we include as main independent variable the 3-month moving average of z-scored mystery shopping performance calculated as the average of all z-scored mystery shopping scores collected for a site in the current, the last, and second to last month. In columns 5 to 10, we include as main independent variable the 6-month, 9-month, or 12-month moving average of the z-scored mystery shopping performance. We calculate these variables analogously to the 3-month moving average (i.e. starting with the current month). In all specifications, we control for the natural logarithm of the monthly hours worked in a site. The odd columns refer to simple linear regressions. The even columns refer to regressions where we include site fixed effects and month fixed effects. Coefficient standard errors are clustered at the site level and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01. We include all site-months for which we have data on the dependent variable, and the independent variables. Furthermore, we drop singleton observations to avoid over-stating statistical significance (Correia 2015).

Table 15: Firm 3 - Regressing sales on lagged mystery shopping scores

	Ln(monthly sales)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Standardized mystery shopping score	-0.019***	-0.008***								
	(0.003)	(0.002)								
1-month lag of standardized mystery shopping score	-0.010***	-0.004**								
	(0.003)	(0.002)								
2-month lag of standardized mystery shopping score	-0.012***	-0.003*								
	(0.003)	(0.002)								
3-month moving average of standardized mystery shopping score			-0.041***	-0.014***						
			(0.007)	(0.004)						
6-month moving average of standardized mystery shopping score					-0.033***	-0.017**				
					(0.012)	(0.007)				
9-month moving average of standardized mystery shopping score							-0.023	-0.014		
							(0.016)	(0.009)		
12-month moving average of standardized mystery shopping score									-0.039**	-0.006
									(0.019)	(0.011)
Ln(monthly hours worked)	0.984***	0.546***	0.984***	0.545***	0.989***	0.518***	0.991***	0.511***	0.986***	0.508***
	(0.013)	(0.100)	(0.013)	(0.100)	(0.013)	(0.102)	(0.013)	(0.110)	(0.013)	(0.120)
Month fixed effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Site fixed effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Clustered at	Site	Site	Site	Site	Site	Site	Site	Site	Site	Site
Number of clusters	241	241	241	241	241	241	241	241	241	241
Observations	7,550	7,550	7,550	7,550	6,777	6,777	6,009	6,009	5,258	5,258
Adjusted R-squared	0.927	0.976	0.927	0.976	0.931	0.980	0.932	0.982	0.933	0.981

Notes: This table reports results of a regression of sales on lagged mystery shopping performance. The data are at the site-month level. The dependent variable is the natural logarithm of monthly sales in Euros. In columns 1 and 2, we include as main independent variable the z-scored mystery shopping performance of the current month, the last month, and the second to last month. We average mystery shopping scores prior to z-scoring if there were multiple mystery shopping visits in a site-month. In column 3 and 4, we include as main independent variable the 3-month moving average of z-scored mystery shopping performance calculated as the average of all z-scored mystery shopping scores collected for a site in the current, the last, and second to last month. In columns 5 to 10, we include as main independent variable the 6-month, 9-month, or 12-month moving average of the z-scored mystery shopping performance. We calculate these variables analogously to the 3-month moving average (i.e. starting with the current month). In all specifications, we control for the natural logarithm of the monthly hours worked in a site. The odd columns refer to simple linear regressions. The even columns refer to regressions where we include site fixed effects and month fixed effects. Coefficient standard errors are clustered at the site level and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01. We include all site-months for which we have data on the dependent variable, and the independent variables. Furthermore, we drop singleton observations to avoid over-stating statistical significance (Correia 2015).

Table 16: Firm 2 - Bonus payments with imputed mystery shopping scores (Heckman)

Imputation method	*Mean*		*Linear Regression*		*Elastic Net*	
	(1)	(2)	(3)	(4)	(5)	(6)
Linear regression	*Ln(bonus in €)*					
Lagged mystery shopping score	0.019		0.021		0.021	
	(0.015)		(0.014)		(0.016)	
Lagged ln(sales)	-0.135		-0.140		-0.140	
	(0.086)		(0.086)		(0.086)	
Lagged ln(sales) -ln(sales target)	2.467***	2.295***	2.491***	2.295***	2.478***	2.292***
	(0.188)	(0.185)	(0.186)	(0.185)	(0.186)	(0.185)
Lagged ln(personnel costs)	0.200**	0.044	0.198**	0.044	0.206**	0.044
	(0.098)	(0.040)	(0.097)	(0.040)	(0.100)	(0.040)
Lagged ln(personnel costs) -ln(personnel costs target)	-0.775***	-0.629***	-0.783***	-0.629***	-0.785***	-0.629***
	(0.088)	(0.073)	(0.087)	(0.073)	(0.089)	(0.073)
Ln(supervisor wage)	0.443***	0.428***	0.446***	0.428***	0.444***	0.428***
	(0.112)	(0.110)	(0.111)	(0.110)	(0.110)	(0.109)
Lagged mystery shopping score missing	0.028		0.027		0.029	
	(0.026)		(0.025)		(0.026)	
Probit	*Received bonus (=1 if yes)*					
Lagged mystery shopping score	0.210***	0.213***	0.206***	0.209***	0.198***	0.201***
	(0.012)	(0.011)	(0.013)	(0.011)	(0.012)	(0.010)
Lagged ln(sales)	-0.288		-0.343		-0.294	
	(0.258)		(0.258)		(0.238)	
Lagged ln(sales) -ln(sales target)	1.126***	0.726***	1.421***	0.937***	1.197***	0.781***
	(0.369)	(0.235)	(0.373)	(0.237)	(0.366)	(0.238)
Lagged ln(personnel costs)	0.536*	0.245***	0.469*	0.121	0.532**	0.243***
	(0.275)	(0.094)	(0.279)	(0.097)	(0.251)	(0.088)
Lagged ln(personnel costs) -ln(personnel costs target)	-0.645**	-0.387*	-0.722**	-0.393*	-0.673**	-0.412**
	(0.293)	(0.209)	(0.305)	(0.209)	(0.283)	(0.208)
Ln(supervisor wage)	0.012		0.037		0.026	
	(0.252)		(0.250)		(0.257)	
Lagged mystery shopping score missing	-0.352***	-0.335***	-0.347***	-0.330***	-0.298***	-0.286***
	(0.075)	(0.077)	(0.077)	(0.079)	(0.082)	(0.082)
Clustered at	Site	Site	Site	Site	Site	Site
Number of clusters	169	169	169	169	169	169
Observations	3,036	3,036	3,036	3,036	3,036	3,036
Complete observations in linear reg.	2,312	2,312	2,312	2,312	2,312	2,312
Complete observations in probit	3,036	3,036	3,036	3,036	3,036	3,036

Notes: This table replicates table 4, columns 2 and 3. However, we impute missing mystery shopping scores with three different predictions. In columns 1 and 2, we impute missing mystery shopping scores with the mean mystery shopping score. In columns 3 and 4, we impute missing mystery shopping scores with predictions from a linear regression where we include as explanatory variables the independent variables used in columns 1,3, and 5 as well as the natural logarithm of monthly hours worked, the natural logarithm of monthly headcount, site fixed effects and month fixed effects. In column 5 and 6, we impute missing mystery shopping scores with predictions from a linear regression chosen to be optimal based on an cross-validated elastic net model featuring the same explanatory variables we use for imputing in columns 3 and 4 as well as their interactions. We drop the first month with potentially usable data (March 2012, recall that variables are lagged by 2 months) because STATA's optimization algorithm did not converge when including this month. Coefficient standard errors are clustered at the site level and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table 17: Firm 3 - Bonus payments with minimum mystery shopping scores (Heckman)

	(1)	(2)
Linear regression	*Ln(bonus in €)*	
Lagged minimum mystery shopping score	0.024***	0.021**
	(0.009)	(0.009)
Lagged ln(sales)-ln(sales target)	0.711***	0.464***
	(0.180)	(0.028)
Lagged ln(sales)	0.508***	0.428**
	(0.170)	(0.175)
Lagged ln(personnel costs)	-0.199	
	(0.196)	
Lagged ln(personnel costs)-ln(personnel costs target)	-0.236	
	(0.202)	
Ln(supervisor wage)	-0.045	
	(0.164)	
Probit	*Received bonus (=1 if yes)*	
Lagged minimum mystery shopping score	0.083***	0.081***
	(0.011)	(0.011)
Lagged ln(sales)	-0.155	
	(0.282)	
Lagged ln(sales)-ln(sales target)	0.949***	1.100***
	(0.355)	(0.344)
Lagged ln(personnel costs)	-0.040	
	(0.311)	
Lagged ln(personnel costs)-ln(personnel costs target)	0.922***	0.344
	(0.351)	(0.254)
Ln(supervisor wage)	0.722***	0.456**
	(0.224)	(0.181)
Clustered at	Site	Site
Number of clusters	241	241
Observations	1,805	1,805
Complete observations in linear regression	975	975
Complete observations in probit	1,805	1,805

Notes: This table replicates table 4, columns 5 and 6. The only change we make is to use as independent variable the quarter-minimum mystery shopping score instead of the quarter-average mystery shopping score. Coefficient standard errors are clustered at the site level and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

B Appendix - Mystery shopping questionnaires

The questionnaire presented first is the mystery shopping questionnaire used by Firm 1 in April 2016. The questionnaire presented second is the mystery shopping questionnaire used by Firm 2 in February 2013. There are minor scoring changes during the observation period but questionnaires remain very similar to the ones below.

FIRM NAME

MS Agency Name

Firm Name

Mystery Questionnaire

Test period 04/2016

1. Info

1.1	**Address of firm name store:** **Street / House Number:**	_____
	ZIP CODE / Location:	_____
1.2	**When did you visit the store of firm name?**	
	☐ Monday ☐ Tuesday ☐ Wednesday ☐ Thursday ☐ Friday ☐ Saturday	
	Test shopping date:	_____._____.2016
	Time:	until ____.____ am/pm
	ⓘ Specify time according to receipt.	
1.3	**Name of test shopper:**	_____
1.4	**Gender:**	☐ Male ☐ Female
1.5	**Age:**	_____ years
1.6	**Receipt no.:**	_____
1.7	**Cash box no.:**	_____
1.8	**Purchase price** **Incl. VAT according to receipt:**	_____ €
1.9	**Item(s) purchased:**	☐ Product A ☐ Product B ☐ I had to buy an alternative product
1.10	**Gender of the employee:**	☐ Male ☐ Female
1.11	**Hair length:**	☐ no hair ☐ short ☐ medium ☐ long ☐ not visible, the employee was wearing a head covering
1.12	**Hair color:**	☐ blonde ☐ black ☐ brown ☐ red ☐ white/grey ☐ not applicable as the employee had no hair ☐ not visible, the employee was wearing a head covering

10 pts. possible

© MS Agency Name 2016 – Firm Name

2. Test Execution

Stage directions:
You visit the respective firm name store and perform the test using the questionnaire and pay attention to the points described there. Pay in cash and mismatch the required amount, so you give the employee ample opportunity to make you an additional offer.
Afterwards, you immediately fill out the questionnaire, either in paper form or mobile in the online tool.

2.1 **Please evaluate the overall visual impression of the firm name store in terms of <u>cleanliness</u> according to the examples in the briefing.**

50 pts. possible
8, 9, 10 = max pts.

1 = bad condition / dirty *10 = perfect condition / clean and tidy*

① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩

2.2 **POSITIVE and/or NEGATIVE peculiarities and comments on the overall impression of "cleanliness" at the firm name store visited (<u>mandatory for all ratings below 8</u>):**

_____

_____

2.3 **Please evaluate the <u>presentation of goods</u> in the firm name store according to the examples in the briefing.**

25 pts. possible
8, 9, 10 = max pts.

1 = poor presentation of goods *10 = perfect presentation of goods*

① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩

2.4 **POSITIVE and/or NEGATIVE peculiarities and comments on the overall impression of "presentation of goods" in the visited firm name store (<u>mandatory for all ratings below 8</u>):**

_____

_____

2.5 **Is there at least one employee in the customer room so that she can respond when customers come in?**

5 pts. possible

☐ Yes

☐ No

2.6 **Did all employees wear work wear?**

20 pts. possible

You can recognize the work wear by the firm name logo.

☐ Yes, all employees wore firm name work wear.

☐ No, not all or no employees wore firm name work wear.

© MS Agency Name 2016 – Firm Name

Stage directions:
When the store employee makes you a concrete additional offer, buy a small coffee / tea (alternatively).
When answering the question about "in-store consumption", please make sure that the firm name store has a seating area. The seating areas are shown in the briefing.

2.7 **Where you greeted in a friendly manner at the checkout?**

A "hello" or a regional expression such as "Moin" or "Grüß Gott" also count as a greeting.

☐ Yes
☐ No

10 pts. possible

2.8 **Did the employee actively make you an additional offer to your purchase?**

The additional offer must be concrete (e.g. a coffee, a coke or another concrete product etc.), see briefing.

☐ Yes
☐ No

2 pts. possible

2.9 **Were you offered a firm name loyalty card by the employee?**

☐ Yes
☐ No

2 pts. possible

2.10 **Before the payment process, were you asked by the employee whether you would like to "dine-in" or "take-away"?**

Asking for one of the two variants (e.g., "take-away?") is sufficient. See briefing!

☐ Yes
☐ No
☐ No, but there was no seating area in the store

8 pts. possible

2.11 **Were you given a friendly farewell or was your farewell returned by the employee?**

A farewell can also be a "goodbye" or a regional expression such as "Servus" or "Ade".

☐ Yes
☐ No

10 pts. possible

2.12 **How would you rate the speed with which you were served at the checkout?**

Note -according to the briefing- that there may be waiting times that the employee cannot influence if there is a high number of guests and only one cash register.

1 = other things were more important, so I had to wait longer 10 = It couldn´t be faster

① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩

10 pts. possible
8, 9, 10 = max pts.

2.13 **POSITIVE and/or NEGATIVE peculiarities and comments on employee behavior (mandatory for all ratings below 8):**

_____

_____

_____

© MS Agency Name 2016 – Firm Name

Report: Firm Name - Checks 2013

Store: Store ID **Check:** 237
Store Address
Time: Wednesday, 13. Feb 2013 / 10:30 A.M. - 10:36 A.M.

	Category Overview		
1	Cleanliness	100.00%	9/9
2	Personnel	91.30%	21/23
3	Product	92.86%	13/14
	Aggregate	**93.48%**	**43/46**

0	**Information**	
0.1.1	First Name of Shopper	John
0.1.2	Surname of Shopper	Doe
0.2.1	Day of Visit	Wednesday
0.2.2	Time Window of Visit	8 - 12 A.M.
0.3.1	Name of Serving Store Employee	Jane Doe
0.4.1	Monthly Competence Question	Are "Berliners" raised or danish pastry?

1	**Cleanliness**			
1.1.1	The merchandise is not handled with bare hands.		Applies	1/1
1.2.1	The glasses of the counter are clean / not blurred.		Applies	1/1
1.3.1	The counter plates are clean / not blurred (only the black plates are to be evaluated).		Applies	1/1
1.4.1	The area behind the counter is clean.		Applies	1/1
1.5.1	The coffee / beverage area is clean and appears tidy.		Applies	1/1
1.6.1	Cleanliness in the eating area.		Applies	3/3
1.7.1	The tables in the eating area are clear.		Applies	1/1
	Overall	**100.00%**		**9/9**

2	**Personnel**			
2.1.1	The sales personnel is available.		Applies	3/3
2.2.1	Greeting with welcome formula.		Applies	1/1
2.3.1	End of interaction with farewell formula.		Applies	1/1
2.4.1	A "Thank you".		Applies	3/3
2.5.1	Friendliness.		Applies	1/1
2.7.1	An additional product was offered (No need that a specific product is offered!).		Applies	3/3
2.8.1	A sample is available.		Applies	1/1
2.9.1	A lye confectionary is offered to each child present in the store.		Applies	1/1
2.10.1	The bread of the month is available.		Applies	1/1
2.11.1	The product of the day is available.		Applies	1/1
2.12.1	Poster ads are in place (all posters are filled with ads and there is a small stand-up display on the counter).		Applies	1/1
2.13.1	Wrapping stickers are used (Anno 1887 or Masterpiece).		Applies	0/1
2.14.1	All lights in the store are working.		Applies	1/1
2.15.1	For 10 randomly chosen products, the label and the price are correct and clearly assigned to the product.		Applies	1/1
2.16.1	The merchandise in the counter is presented appealingly.		Applies	1/1
2.17.1	The workwear is complete.		Applies	1/1
2.18.1	The workwear of staff appears neat / clean.		Applies	0/1
	Overall	**91.30%**		**21/23**

3	**Product**			
3.1.1	The roll tastes fresh and crunchy.		Applies	3/3
3.2.1	The sample tastes fresh.		Applies	1/1
3.3.1	Seven kinds of bread/baguette are available.		Applies	1/1
3.4.1	Five different kinds of rolls are available.		Applies	1/1
3.5.1	Six different kinds of sweet baked goods (cakes, buns etc.) are available.		Applies	0/1
3.6.1	Lye pretzels are available.		Applies	1/1
3.7.1	Snacks: There are at least two pieces of the snack of the month available.		Applies	1/1
3.7.2	Snacks: There are at least four different sandwiches (incl. the snack of the month) available in the cooling area.		Applies	1/1

3.7.3	Snacks: There are at least three different bakery snacks (e.g. cheese roll, cheese pretzl or pizza slice) available.	Applies	1/1
3.8.1	The visual quality of rolls, grain roles and rye roles is decent.	Applies	1/1
3.9.1	Sandwiches appear fresh and delicious.	Applies	1/1
3.10.1	There is confectionery merchandise or other merchandise on display in the cooling area and it looks fresh and delicious.	Applies	1/1
	Overall **92.86%**		**13/14**

4	**Comments**	
4.1.2	Monthly competence question: Are "Berliners" raised or danish pastry?	"Berliners are raised pastry."
4.2.1	Comments on the test visit in general. Please note again the name of the store employee who served you. If merchandise was not available, please note which merchandise was not available anymore.	Mrs. Doe served competently, made an additional offering and offered a sample. She answered the competence question in a friendly manner. A sizeable and nice seating area is available and appealingly designed.

C Appendix - Derivation of the analytical results

Here we explicitly derive our main analytical results presented in sections 2.2.2 and 2.3.1: the expressions for the firm- and site-preferred service levels (equations (1) and (2)), lower sensitivity of the firm-preferred service level to resource budget ($\frac{ds_i^*}{db_i} < \frac{ds_i^{**}}{db_i}$), and how service effort complementarities captured in the complementarity parameter ρ affect firm-preferred service effort sensitivity to site resource budget, $\frac{ds_i^*}{db_i}$.

The firm's optimization problem is

$$\max_{s_i} \sum_{i=1}^{n} LV_i = r(s_1, s_2, ..., s_i, ..., s_n) \cdot \sum_{i=1}^{n} g(p_i, s_i),$$

the first-order condition to which is (skipping arguments for brevity)

$$\frac{\partial r}{\partial s_i} \cdot \sum g + r \cdot \left[\frac{\partial g}{\partial p_i} \cdot \underbrace{\frac{dp_i}{ds_i}}_{=-1} + \frac{\partial g}{\partial s_i} \right] = \frac{\partial r}{\partial s_i} \cdot \sum g - r \cdot \left[\frac{\partial g}{\partial p_i} - \frac{\partial g}{\partial s_i} \right] = 0$$

Dividing the above by $r \cdot \sum g$ gives

$$\frac{\partial r/r}{\partial s_i} = \left[\frac{\partial g}{\partial p_i} - \frac{\partial g}{\partial s_i} \right] \cdot \frac{1}{\sum g} = \left[\frac{\partial g/g}{\partial p_i} - \frac{\partial g/g}{\partial s_i} \right] \cdot \frac{g}{\sum g}$$

Noting that $\partial x/x = \partial \ln(x)$ and $\frac{g}{\sum g} = \phi$, the above can be rewritten as equation (1). Equation (2) is derived similarly.

To derive the result in (3), $\frac{ds_i^*}{db_i} < \frac{ds_i^{**}}{db_i}$, compute the implicit derivatives $\frac{ds_i^*}{db_i}$ from (1) and $\frac{ds_i^{**}}{db_i}$ from (2). Starting with (1), which we rewrite for convenience as $R - G \cdot \phi_i = 0$, where $R = \frac{\partial \ln(r(..., s_i^*, ...))}{\partial s_i}$ and $G = \left[\frac{\partial \ln(g(p_i^*, s_i^*))}{\partial p_i} - \frac{\partial \ln(g(p_i^*, s_i^*))}{\partial s_i} \right]$,

$$\frac{ds_i^*}{db_i} = -\frac{\frac{\partial (R - G \cdot \phi_i)}{\partial b_i}}{\frac{\partial (R - G \cdot \phi_i)}{\partial s_i^*}} = \frac{\frac{\partial G}{\partial b_i} \cdot \phi_i + \frac{\partial \phi_i}{\partial b_i} \cdot G}{\frac{\partial R}{\partial s_i^*} - \frac{\partial G}{\partial s_i^*} \cdot \phi_i - \frac{\phi_i}{\partial s_i^*} \cdot G} \approx \frac{\frac{\partial G}{\partial b_i} \cdot \phi_i}{\frac{\partial R}{\partial s_i^*} - \frac{\partial G}{\partial s_i^*} \cdot \phi_i} = \frac{\frac{\partial G}{\partial b_i}}{\frac{1}{\phi_i} \frac{\partial R}{\partial s_i^*} - \frac{\partial G}{\partial s_i^*}}$$

In deriving the above expression, we ignore the effects of an individual site's resource budget and service level on its share in total sales, that is, we set $\frac{d\phi_i}{ds_i} = 0$ and $\frac{d\phi_i}{db_i} = 0$. This is a permissible approximation when the number of sites is large, as is the case with all our study firms. Proceeding with (2), which, using the same notations as above,

can be rewritten as $R - G = 0$,

$$\frac{ds_i^{**}}{db_i} = -\frac{\frac{\partial(R-G)}{\partial b_i}}{\frac{\partial(R-G)}{\partial s_i^{**}}} = \frac{\frac{\partial G}{\partial b_i}}{\frac{\partial R}{\partial s_i^{**}} - \frac{\partial G}{\partial s_i^{**}}}$$

Comparing the above expressions for $\frac{ds_i^{*}}{db_i}$ and $\frac{ds_i^{**}}{db_i}$ and noting that $\frac{1}{\phi_i} >> 1$ when the number of sites is large, proves (3).

To see how the strength of service effort complementarities ρ affect optimal service effort sensitivity to resource budget, consider a second-order Taylor-series approximation of the reputation factor around the mean service effort (Friebel et al. (2021) provide a full derivation of this result):

$$r(s_1, ..., s_i, ..., s_n) = a \cdot \left(\sum_{i=1}^{n} s_i^{\rho} \right)^{\frac{1}{\rho}} \approx a \cdot n \cdot \left(\bar{s} + \frac{1}{2}(\rho - 1)\frac{var(s)}{\bar{s}} \right)$$

Focusing on the term $\frac{1}{2}(\rho - 1)\frac{var(s)}{\bar{s}}$, a mean-preserving increase in the service effort variation across sites is detrimental to the firm's reputation with efforts are complementary, that is, when $\rho < 1$. It follows that strong service effort complementarities should suppress the optimal service effort variation with resource budget across sites.

Lastly to derive the result stated in Section 2.3.1 that $r(s_1, ..., s_i, ..., s_n) = a \cdot \min(s_1, ..., s_i, ..., s_n)$ when complementarities are extreme, $\rho \to -\infty$, compute the limit

$$\lim_{\rho \to -\infty} a \cdot \left(\sum_{i=1}^{n} s_i^{\rho} \right)^{\frac{1}{\rho}}$$

Choosing $s_m = \min(s_1, ..., s_i, ..., s_n)$ and factorizing in terms of s_m, we obtain

$$a \cdot s_m \cdot \lim_{\rho \to -\infty} \left(\sum_{j \neq m}^{n-1} \underbrace{\left[\frac{s_j}{s_m} \right]^{\rho}}_{\to 0} + 1 \right)^{\frac{1}{\rho}} = a \cdot s_m.$$

When the entire firm's reputation is determined by the worst service level found among its sites, s_m, the profit maximizing s_m should be chosen as the service standard, meeting which is essential but exceeding is not necessary.