Journal of Retailing 95 (3, 2019) 47–62

# Do Mystery Shoppers Really Predict Customer Satisfaction and Sales Performance?

Gerald Blessing [a,*], Martin Natter [b]

[a] *ESB Business School, Reutlingen University, Germany*
[b] *Department of Business Administration, University of Zurich, Switzerland*

Available online 16 May 2019

**Abstract**

Mystery shopping (MS) is a widely used tool to monitor the quality of service and personal selling. In consultative retail settings, assessments of mystery shoppers are supposed to capture the most relevant aspects of salespeople's service and sales behavior. Given the important conclusions drawn by managers from MS results, the standard assumption seems to be that assessments of mystery shoppers are strongly related to customer satisfaction and sales performance. However, surprisingly scant empirical evidence supports this assumption. We test the relationship between MS assessments and customer evaluations and sales performance with large-scale data from three service retail chains. Surprisingly, we do not find a substantial correlation. The results show that mystery shoppers are not good proxies for real customers. While MS assessments are not related to sales, our findings confirm the established correlation between customer satisfaction measurements and sales results.


*Keywords:* Mystery shopping; Personal selling; Sales management; Customer satisfaction

## Introduction

Mystery shopping (MS) is a participant observation method that many companies and public organizations use to measure the quality of service delivery (Wilson 2001). Commercially, MS is very successful: in 2016, the Mystery Shopping Professional Association (MSPA 2018a) set worldwide spending for MS at $2 billion, with the United States accounting for half the market and Europe for approximately half a billion dollars. Overall, the MSPA (2018b) estimates that there are 1.5 million mystery shoppers worldwide. Although more detailed data on the prevalence of MS across industries are not publicly available, finance, telecommunications, retail, leisure/travel, hospitality, and motor dealerships are considered the main areas of use (Van der Wiele, Hesselink, and Van Iwaarden 2005), and many companies in these sectors run MS programs on a regular basis.

In our research, we refer to the use of MS in consultative retail settings, in which skilled salespeople determine the perceived service quality and the sales results to a great extent (Grewal, Levy, and Marshall 2002). In these settings (consumer-durables retailers, service retailers, car dealers), managers want to know how their salespeople are perceived and how they behave. They use MS to measure the quality of the personal selling and rely on MS results to make managerial decisions, to benchmark retail stores, to set up sales training programs, and to evaluate and reward their sales staff. By doing so, they implicitly assume that mystery shoppers provide accurate assessments of the personal selling and that these assessments reflect the outcomes of sales encounters—in particular, customer satisfaction and sales performance. Conventional wisdom and the service-profit chain framework support the relationship among salespeople's behavior, customer satisfaction, and sales results (Anderson and Mittal 2000; Heskett et al. 1994); however, despite the high managerial relevance, this relationship has not yet been tested.

Our aim is to provide insights into the informative value of MS data. Our research question is whether MS assessments are related to customer satisfaction and objective sales performance. With our research, we intend to provide valuable information for managers who need to know whether MS data are a reliable basis for decision making. Our research is also relevant for the acceptance of MS among salespeople and contributes to sales force research. Until now, the measurement of the determinants

* Corresponding author. Tel.: +49 7121 271 1432.
*E-mail addresses:* Gerald.Blessing@Reutlingen-University.de (G. Blessing), martin.natter@business.uzh.ch (M. Natter).

https://doi.org/10.1016/j.jretai.2019.04.001

of sales performance has been based on self-reports, managerial judgments, and customer evaluations (Verbeke, Dietz, and Verwaal 2011), despite the serious shortcoming of these techniques (Jaramillo, Carrillat, and Locander 2005; Levy and Sharma 1993). Evidence of a significant relationship between MS judgments and sales outcomes would confirm the use of MS as an alternative tool to measure salespeople's attributes, as in the work of Price, Arnould, and Deibler (1995).

We begin our article with a brief introduction of the MS method and then propose a conceptual framework that relates the assessments of salespeople's attributes to the satisfaction with salespeople and to sales performance. We then test our framework using large-scale MS data from three service retail companies. We find that the assessments of mystery shoppers are not related to customer satisfaction. In addition, customer evaluations are predictive of sales performance, while MS assessments are not. Drawing on these results, we discuss the managerial implications and propose an agenda for further MS research.

## Mystery Shopping

Mystery shoppers, often called anonymous, silent, or secret shoppers, visit service points or stores, pretend to be normal customers, observe the process of service delivery, and, immediately after the service interaction, record their observations on different aspects of the service experience in a detailed questionnaire (Finn and Kayandé 1999). Mystery shoppers assess objective aspects of a service encounter (e.g., did the salesperson ask a closing question?) and can also evaluate subjective aspects (e.g., friendliness, competence), which are usually surveyed from real customers (Finn and Kayandé 1999). Mystery shoppers attentively monitor the process of service delivery and thus can evaluate very specific aspects of the service interaction, in contrast with normal customers who mostly do not recall particular details of a service experience. Mystery shoppers' measurement of service quality is also supposed to be more objective than managerial judgments, employees' self-reports, or customer evaluations (Wilson 2001). Another advantage over customer surveys is the flexibility of the MS tool, especially in settings in which it is difficult to collect customer responses.

Some evidence suggests that MS programs can lead to higher service performance (Van der Wiele, Hesselink, and Van Iwaarden 2005). In Wilson's (2001) study, practitioners reported that MS has at least a short-term impact on service standards, even though at a later stage the effects tend to reach "a plateau of no further improvement." For these reasons, MS has become an important tool not only to measure the quality of service provision but also to develop service employees, to benchmark service performance, and as a basis for managerial decisions and training programs (Wilson 2001).

In general, MS data are surveyed at the store level, though observations at the individual level are possible if salespeople explicitly agree to be personally evaluated. It is common practice for MS agencies to aggregate the ratings collected by two to four mystery shoppers at the store level and use the mean ratings to calculate performance indicators for different aspects of the sales process (e.g., different stages of a sales encounter) or for different selling skills (e.g., relationship quality, consultation quality). Often, these performance indicators are then used to calculate overall performance indices.

A bulk of the research in specialized journals provides application examples of MS from different industries (e.g., Calvert 2005; Erstad 1998; Mattson 2011; Peterman and Young 2015; Van der Wiele, Hesselink, and Van Iwaarden 2005; Xu and He 2014). However, publications in academic marketing journals are scarce. So far, the articles published approximately 20 years ago by Finn (2001), Finn and Kayandé (1999), Morrison, Colman, and Preston (1997), and Wilson (1998a, 1998b, 2001) are among the most cited on MS. Wilson (1998a, 1998b, 2001) interviewed 10 senior managers to uncover the views of practitioners on the method and how they use MS in practice. Morrison, Colman, and Preston (1997) outline how the biases associated with the encoding, storage, and retrieval of information by mystery shoppers can affect the reliability of MS observations. Finn (2001) and Finn and Kayandé (1999) investigate the psychometric quality of MS. They find that MS assessments are more cost-effective than customer surveys but claim that the industry practice of using two–four mystery visits to measure service quality and personal selling is insufficient to provide representative results.

## Conceptual Framework and Hypotheses

Although an increasing number of customers decide on the purchase process before they enter a retail shop, industry reports suggest that between 40 and 70% of customers are still open to persuasion and make their buying decisions after they enter a store (Leibowotz 2010; Neff 2008). Especially in consultative retail settings, salespeople play a decisive role for customer behavior and the business success of retail shops (Brady and Cronin 2001; Macintosh and Lockshin 1997; Sweeney, Soutar, and Johnson 1997; Westbrook 1981).

Salespeople draw on social and interpersonal skills to meet customers' emotional needs and make use of task-related competencies to help customers achieve their purchase goals (Brexendorf et al. 2010; Van Dolen et al. 2002). Customers' perceptions of salespeople's attributes and behavior influence their satisfaction with the salesperson and the sales encounter. The relationship between perceptions of salespeople attributes and customer satisfaction is conceptually supported by the expectancy–disconfirmation paradigm and the theory of planned behavior and is empirically documented by several studies in the retailing sector (Ailawadi et al. 2014; Brexendorf et al. 2010; Crosby, Evans, and Cowles 1990; Homburg, Müller, and Klarmann 2011; Hunneman, Verhoef, and Sloot 2015; Swan and Oliver 1991; Van Dolen et al. 2002). Accordingly, we propose the following:

**H1a.** Customer perceptions of salespeople attributes are correlated with customer satisfaction with a salesperson.

Research based on the service–profit chain framework (Heskett et al. 1994) demonstrates that service quality affects customer satisfaction and that customer satisfaction, in turn,

is a key predictor of purchase intentions (Anderson and Mittal 2000; Carrillat, Jaramillo, and Mulki 2009; Grewal and Sharma 1991). Empirical studies in the retail sector also provide evidence of a positive influence of customer satisfaction on sales. Babin, Babin, and Boles (1999) report a positive relationship between the attitude toward a retail salesperson and purchase intentions, mediated by the attitude toward the retailer. Similarly, with data from other sectors, Ahearne, Mathieu, and Rapp (2005), Brady and Cronin (2001), and Homburg, Müller, and Klarmann (2011) report a positive effect of a customer's attitude toward the salesperson on customer satisfaction and behavioral outcomes such as sales performance. Gomez, McLaughlin, and Wittink (2004) show that even in the food retail sector, customer service perceptions are highly influenced by retail employees and that service perceptions are positively related to store sales performance. Together, these findings indicate that satisfied customers purchase more from salespeople with whom they are satisfied. Thus:

**H1b.** Customer satisfaction with a salesperson is positively correlated with the objective sales performance of the salesperson.

Evidence indicates that customer satisfaction only partly mediates the impact of retail store attributes on sales performance (Ailawadi et al. 2014; Hunneman, Verhoef, and Sloot 2015). Because persuasion is at the heart of the sales role, salespeople may use a route of persuasion (Babin, Babin, and Boles 1999) that directs customers' buying behavior without them being aware of being influenced. For example, by selecting and disseminating relevant information, as well as by using sales tactics, salespeople can influence customers' buying decisions independent of customer satisfaction (Gabler et al. 2017; McFarland, Challagalla, and Shervani 2006; Plouffe, Bolander, and Cote 2014). Thus:

**H1c.** Customer perceptions of salespeople attributes are correlated with sales performance.

In general, customer surveys measure attribute level and overall customer satisfaction. The question arises whether mystery shopper assessments can serve as substitutes for customer evaluations: are mystery shoppers good proxies of real customers and representative of the customer population? Conventional marketing wisdom supports this assumption. On the one hand, MS is a customer-oriented method of service measurement, and mystery shoppers are supposed to monitor the service delivery through the eyes of a customer. Therefore, mystery shoppers experience the sales interaction like a real customer, and thus it is fair to assume that their assessment of a salesperson mirrors the evaluations of real customers. In terms of objective attributes, the information mystery shoppers provide should be consistent with respective observations of real customers. Their assessments may even be more reliable because mystery shoppers attentively observe specific details of the sales encounter. On the other hand, many factors may influence and bias subjective assessments of salespeople, including the expectations, attitudes, involvement, and product experience of the assessor (Morrison, Colman, and Preston 1997). Therefore, professional mystery research agencies try to select mystery shoppers who are most representative of the real customer population. We, therefore, assume that subjective assessments of mystery shoppers and real customers are consistent. Several empirical studies substantiate this assumption: Wilson and Gutmann (1998) and Finn and Kayandé (1999) report significant correlations between overall customer satisfaction scores and average mystery scores. More recently, Hoekstra, Ammeraal, and Leeflang (2014) observed that the satisfaction ratings of real customers are well reflected by mystery callers' judgments. We therefore assume that mystery shoppers are able to provide accurate measures of customer evaluations. In line with H1a, we also assume that mystery shoppers' attribute-level perceptions are the main drivers of their overall satisfaction with the salesperson. Thus:

**H2a.** Mystery shoppers' and real customers' assessments of salespeople attributes are positively correlated.

**H2b.** Mystery shoppers' and real customers' satisfaction with a salesperson are positively correlated.

Moreover, we assume that the chain of effects of salespeople attributes, satisfaction with salespeople, and sales performance, as outlined in H1, holds for mystery shoppers in the same way as for real customers. Thus:

**H3a.** Mystery shoppers' assessments of salespeople attributes are correlated with their satisfaction with a salesperson.

**H3b.** Mystery shoppers' satisfaction with a salesperson is positively correlated with the objective sales performance of the salesperson.

**H3c.** Mystery shoppers' assessments of salespeople attributes are correlated with sales performance.

While our hypotheses might seem straight forward at first glance, several factors may compromise the relationship between MS assessments of salespeople attributes and sales outcomes. First, mystery shoppers may not be representative of the customer population. Even if mystery shoppers are carefully selected to match the profile of normal customers as best as possible, important differences remain. In particular, mystery shoppers are not personally involved as real customers are. They do not depend on the salesperson's assistance to make the best possible buying decision and do not take on any risk of decision making. If customers are experienced, they may also not have the same level of product knowledge as real customers.

Second, the number of mystery visits may also affect the association between MS assessments and sales outcomes. To obtain sufficiently generalizable measures for benchmarking retail stores, Finn (2001) advocates for at least 20 visits per shop, much more than the industry standard, which is two to four visits per observational unit. Finn's recommendation is based on a field study that does not control for the influence of different salespeople. Nonetheless, the number of available observations is a critical aspect of the MS method. However, for several reasons, increasing the number of visits per store to 20 or more is not a viable option for many companies. The

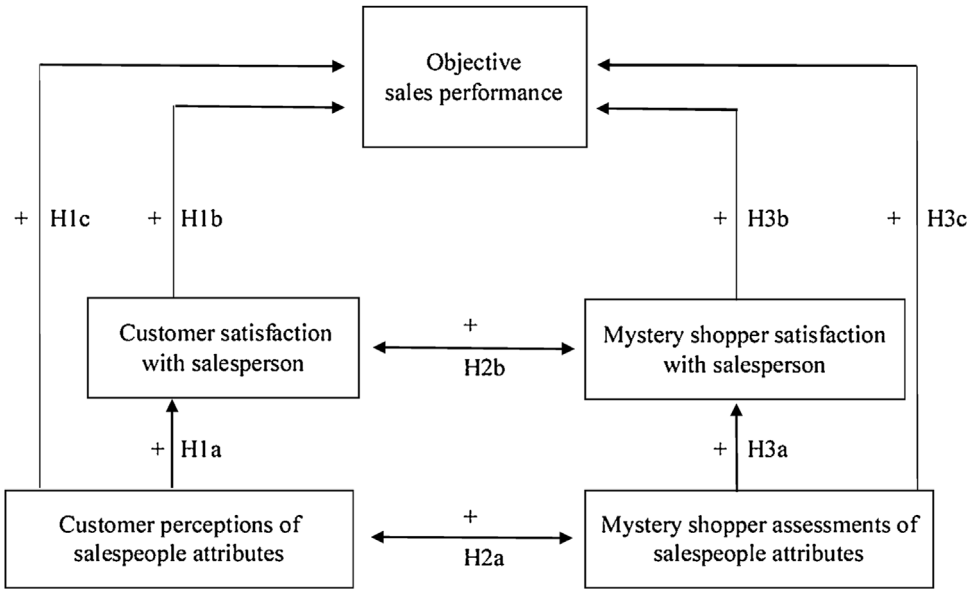


Fig. 1. Conceptual framework.

costs of data collection would significantly rise, and a high number of visits risks compromising the normal course of business, thus conflicting with ethical standards of MS (ESOMAR 2005). In addition, a significant increase in the number of MS visits would likely trigger negative reactions from the salespeople concerned (and from work councils), and the risk of unmasking mystery shoppers—another threat to the reliability of MS observations—would also increase.

Third, the complexity of the MS task also jeopardizes the predictive ability of MS assessments. Many MS surveys are extensive, containing a large number of questions. Processing all this information is extremely demanding and generates the risk of biased observations. For example, mystery shoppers may be tempted to substitute missing perceptions or memory gaps with extrapolations of their overall impression (Wirtz 2000) or, if experiencing difficulties in evaluating attributes of a salesperson (e.g., product expertise), may resort to easy-to-evaluate attributes to manage the evaluation task (Dagger et al. 2013).

Fourth, the choice of sales scenarios and the scaling of salespeople attributes may also affect the accuracy of MS data. The sales scenarios mystery shoppers use should be representative of the salespeople's daily business. At the same time, the scenarios must be sufficiently demanding to uncover the differences between salespeople, even when these differences are subtle, and to avoid possible ceiling effects. The same requirement applies to the scales used in MS surveys. In the absence of a widely accepted scale to measure the quality of personal selling (Finn 2001), MS agencies generally use proprietary measures to operationalize salespeople attributes. To the best of our knowledge, empirically validated service measurement scales such as the SERVQUAL approach are generally not used in MS studies, though there is good rationale to do so (Lowndes and Dawes 2001). Therefore, if proprietary scales do not provide a valid measurement of the salespeople attributes, MS assessments run the risk of not being related to customer satisfaction and sales results.

In summary, significant risks may impede the relationship between MS assessments and customer satisfaction and/or sales performance. Therefore, it is important to test this relationship empirically. Fig. 1, summarizes our hypotheses in a conceptual framework that we use as a guideline for our analysis. Because our primary interest is to examine whether mystery shopper assessments are good predictors of customer satisfaction and sales performance, we focus specifically on H2a, H2b, H3b, and H3c. We use the relationship among customer perceptions, customer satisfaction, and sales performance (H1a–H1c), which is well established in the literature, to cross-validate the MS results.

## Overview of Studies

Three service companies, all operating in the same sector, showed interest in our research and agreed to support our project on MS and sales performance data. In addition, one of the companies provided us with customer satisfaction data that had been surveyed in parallel to MS visits. We use this data set to examine the relationship between the assessments of mystery shoppers and real customers in Study 1. In Study 2, we use MS and sales data from the three companies to test the association between MS data and sales performance.

## Study 1: Relationship between MS Assessments and Evaluations of Real Customers

### *Data*

Our first data set contains MS and customer satisfaction assessments surveyed in 2014 in 204 shops of our partner company. The company sells high-value services with low purchase frequency, long consumption times, and high financial commitment, as is the case with energy, telecom, and insurance products. The shops have the same design, they sell the same products

Table 1
Scale Items in MS and customer survey (data set 1).

| MS survey (n = 611) | Customer survey (n = 21,182) |
|---|---|
| ***Subjective items*** | |
| Salesperson has **expertise**.[a] (M: 4.28, Mdn: 4, SD: .87) | Satisfaction with salesperson's **expertise**.[b] (M: 4.04, Mdn: 4, SD: .91) |
| Salesperson gives **comprehensive information**. (M: 4.22, Mdn: 4, SD: .93) | Satisfaction with **comprehensiveness of information** provided by the salesperson. (M: 3.98, Mdn: 4, SD: .92) |
| How did you perceive the salesperson's **friendliness?** (M: 4.46, Mdn: 5, SD: .77)[c] | Satisfaction with salesperson's **friendliness.** (M: 4.25, Mdn: 4, SD: .88) |
| Salesperson **took my concern seriously.** (M: 4.25, Mdn: 5, SD: 1.0) | Satisfaction with how the salesperson **took your concerns seriously.** (M: 4.11, SD: .88) |
| Salesperson **took enough time** to consult me. (M: 4.28, Mdn: 5, SD: 1.07) | Satisfaction with salesperson's **personal effort.** (M: 4.15, Mdn: 4, SD: .88) |
| ***Objective items*** | |
| Did a salesperson **welcome** you when entering the shop? (M:.65, Mdn: 1, SD:.48)[d] | Did a salesperson **welcome** you when entering the shop? (M: .71, Mdn: 1, SD: .45)[d] |
| Did the salesperson make a **handwritten offer**? (M:.82, Mdn: 1, SD:.38)[d] | Did the salesperson make **handwritten offer?** (M: .46, Mdn: 0, SD: .50)[d] |
| Did the salesperson **handout your documents** in a folder? (M: .44, Mdn: 0, SD:.50)[d] | Did the salesperson **handout your documents** in a folder? (M: .64, Mdn: 1, SD: .48)[d] |
| Did the salesperson make an **additional offer?** (M: .45, Mdn: 0, SD:.50)[d] | Did the salesperson make an **additional offer?** (M: .48, Mdn: 0, SD: .86)[d] |
| Further question items | |
| Salesperson was . . . well informed, provided many information, gave structured explanations, responded to individual needs etc. | **Satisfaction** with **salesperson's consultation service**<br>**Satisfaction** with **shop visit** |

[a] Subjective assessments in MS survey measured on reversed five-point scales (1 = "do not agree at all", 5 = "fully agree").
[b] Subjective assessments in customer survey measured on reversed 5-point scales (1 = "not satisfied", 5 = "extremely satisfied").
[c] Reversed five-point scale (1 = "not friendly", 5 = "very friendly").
[d] Binary scale.

that are exclusively from the service company, and the regular prices for the services are the same. Thus, the salespeople are the primary distinguishing component among the shops. MS data are available from three mystery visits per shop, and customer evaluations derive from 21,182 after-customer contact telephone interviews (mean number of observations per shop: 103.88, SD: 11.98).

### *Measures*

The MS survey contains more than 60 question items. In addition to shop-related information (e.g., type and location of the shop, waiting time), the survey comprises subjective assessments of salespeople attributes (e.g., expertise, friendliness) and detailed observations about the salespeople's behavior during the sales encounter (e.g., if the salesperson asked specific questions). The subjective assessments were measured with five-point Likert-type scales, the observational items with binary scales.

The customer survey contains 21 questions, mostly satisfaction assessments and objective observations of the salesperson but also customer-related information (e.g., private or business customer). Two items measure customers' satisfaction with the salesperson and with the shop visit overall, on five-point-Likert scales, as is the case with all other subjective assessments. Nine salesperson attribute items in the customer and the MS survey are largely identical (see Table 1). Five of these items are subjective attribute assessments, and four items are related to a salesperson's behavior, objectively measured by simple observation (e.g., welcome greeting).

### *Results*

To test H2a, we calculated bivariate correlations between the modes and means of the mystery shoppers' and real customers' ratings on the nine identical scale items across all 204 shops. The correlations between the mystery shopper and customer rating modes are low, with Kendall's tau ranging from –.09 (salesperson offered an additional product) to .22 (handout at exit). Correlation analysis based on mean rating values for the nine scale items yields similar results. Kendall's tau varies between –.03 (salesperson offered an additional product) and .3 (handout at exit). We find that five of the nine correlation coefficients are significant ($p < .05$). However, the correlation levels between the average assessments of mystery shoppers and real customers for almost identical questions are all low. Pearson's and Spearman's correlation coefficients provide similar results. Thus, our data, surveyed with industry standard MS studies and after-contract customer interviews, do not support H2a for four of the nine items. Although, we find five significantly correlated items, the correlation levels are low or even negative, suggesting that, overall, assessments of mystery shoppers are not or only weakly correlated with those of real customers.

In the next step, we test H2b (i.e., whether mystery shoppers' overall assessment of a salesperson is positively correlated with customers' satisfaction with a salesperson). While we have access to data about customers' satisfaction with the salesperson and the shop visit, our MS data do not report the mystery shoppers' explicit overall assessment of a salesperson and shop visits. As an approximation, we take the sum of the individual subjective assessments of each mystery shopper about a salesperson. Analysis of data sets 2 and 3, in which we actually have

Table 2
Customer satisfaction: mixed-effects model estimates (MS and customer data, data set 1).

| Dependent variable: customer satisfaction with salesperson (individual ratings) | | | |
|---|---|---|---|
| Predictor | MS data<br>Estimate (SE) | Customer data<br>Estimate (SE) | Customer & MS data<br>Estimate (SE) |
| Intercept | 4.12 (.01)*** | 4.02 (.01)*** | 4.02 (.01)*** |
| Factor 1 (interpersonal) | .03 (.02) | | .00 (.01) |
| Factor 2 (competence) | −.00 (.02) | | .01 (.01) |
| Handout at exit[m] | −.01 (.02) | | −.01 (.01) |
| Handwritten offer[m] | .01 (.02) | | −.01 (.01) |
| Welcome greeting[m] | .01 (.02) | | −.01 (.01) |
| Additional offer[m] | −.04 (.02)* | | −.00 (.01) |
| Factor 1 (interpersonal) | | .77 (.01)*** | .77 (.01)*** |
| Factor 2 (competence) | | .84 (.01)*** | .84 (.01)*** |
| Handout at exit[d] | | .05 (.01)*** | .05 (.01)*** |
| Handwritten offer[d] | | .01 (.01) | .01 (.01) |
| Welcome greeting[d] | | .07 (.01)*** | .07 (.01)*** |
| Additional offer[d] | | .02 (.01)* | .02 (.01)* |
| $\Delta\chi^2$ | 4.30 (d.f. = 6) | 10,119.41 (d.f. = 6)*** | 10,122.29 (d.f. = 12)*** |

n = 18,127 customer assessments and n = 611 MS assessments on 204 shops.
Dependent variable measured with inversed five-point scale (1 = "not satisfied", 5 = "strongly satisfied").
Model fit by maximum likelihood. Incremental chi-square (log-likelihood) compared to random intercept model.
[d] binary variable. [m] mean value of binary variable across mystery shoppers.
All variables, except binary variables, were standardized by dividing by 2 standard deviations (Gelman 2008).
* $p < .05$.
*** $p < .001$.

these data (though no customer satisfaction data), shows that the mystery shoppers' assessments across all subjective salespeople attributes are highly correlated with their overall assessment of a salesperson. Pearson's correlation between the mean customer satisfaction with the salesperson and this composite measure for the mystery shoppers' satisfaction is r = .12 ($p > .05$). Thus, our results provide no support for H2b.

To test H1a, we analyze the relationship between customer perceptions of salespeople attributes and customer satisfaction with the salesperson. We find that customer satisfaction with the salesperson is highly correlated with the customers' own assessments of the nine salespeople items (supporting H1a). All these correlations are significant ($p < .01$), and Pearson's r varies between .68 and .74 for the subjective items and between .14 and .20 for the objective measures. Thus, the customers' overall satisfaction is associated more with their assessments of subjective salespeople attributes than with their objective observations of salespeople behavior.

After inspecting the correlations between mystery shoppers' and customer-based assessments of salespeople, we examine the multivariate effect of salespeople's attributes on customer satisfaction. Given the high correlations of the five subjective evaluations of both mystery shoppers and real customers, we performed a principal axis factor analysis with varimax rotation with the ratings of each group of assessors. As a result, we obtained two factors, each with eigenvalues greater than 1. One factor represents the salesperson's interpersonal skills (e.g., friendliness, taking customer concerns seriously, taking enough time) and the other the salesperson's competence (e.g., expertise, comprehensiveness of the information provided).

Because the MS and customer assessments of salespeople are nested within shops, we use a hierarchical mixed-effects model (Snijders and Bosker 2012) to regress individual customer satisfaction with a salesperson on salespeople attributes. We estimate separate models to assess the relationship between (1) individual customer satisfaction and individual customer perceptions of salespeople attributes and (2) individual customer satisfaction and mean mystery shopper assessments of salespeople.

In both cases, we use the two skill factors extracted from the customers' and mystery shoppers' attribute ratings as independent variables. To test whether companies that use both data sources could benefit from additional insights, we then estimated the relationship between overall customer satisfaction and both mystery shopper and customer evaluations (see Table 2).

We find that the model based on MS assessments does not provide an improvement of fit over a basic random intercept model. Except for one item, which shows a weak effect (additional offer), all predictors based on MS assessments are non-significant. By contrast, the model based on customer data shows a significantly higher goodness-of-fit than a basic random intercept model. The competence factor and the interpersonal skill factor, based on customer responses, have the strongest effect on overall satisfaction with a salesperson, and except for two items, the effects of all customer measures are highly significant ($p < .001$). Not surprisingly, using MS assessments in addition to customer data does not improve the predictive quality of the model. We find that customer satisfaction with a salesperson is related neither to subjective nor to objective observations of mystery shoppers. Thus, our results for overall customer satisfaction with a salesperson are consistent with the finding that the assessments of mystery shoppers and real customers are not

Table 3
Sales performance: OLS regression estimates (MS and customer data, data set 1).

| Dependent variable: logit of average shop conversion rate. | | | |
|---|---|---|---|
| Predictor | MS data<br>Estimate β (SE) | Customer data<br>Estimate β (SE) | MS and customer data<br>Estimate β (SE) |
| Intercept | −4.60 (.03)*** | −4.60 (.03)*** | −4.60 (.03)*** |
| Factor 1 (interpersonal) | .03 (.07) | | .01 (.06) |
| Factor 2 (competence) | −.10 (.06) | | −.10 (.05) |
| Handout at exit[m] | .17 (.07)* | | .04 (.06) |
| Handwritten offer[m] | −.03 (.07) | | −.04 (.06) |
| Welcome greeting[m] | −.09 (.06) | | −.05 (.06) |
| Additional offer[m] | −.00 (.06) | | −.01 (.05) |
| Factor 1 (competence) | | .25 (.06)*** | .26 (.06)*** |
| Factor 2 (interpersonal) | | −.08 (.06) | −.08 (.06) |
| Handout at exit[m] | | .26 (.06) *** | .25 (.07)*** |
| Handwritten offer[m] | | .03 (.07) | .03 (.08) |
| Welcome greeting[m] | | −.23 (.05) *** | −.22 (.06)*** |
| Additional offer[m] | | −.01 (.07) | .01 (.07) |
| $R^2$ adj. | .02 | .25 | .24 |
| F-statistic | $F_{(6, 166)} = 1.66$ | $F_{(6, 166)} = 10.44$*** | $F_{(12, 160)} = 5.59$*** |

n = 173. All VIF values < 2.
[m] mean values of binary variable. All variables were standardized by dividing by 2 standard deviations.
*$p < .05$.
***$p < .001$.

or only weakly correlated at the attribute level. Overall, as there is no evidence supporting our hypotheses H2a and H2b, we conclude that mystery shopper assessments are not good proxies for real customer evaluations.

## Study 2: Relationship between MS Assessments and Sales Performance

### *Data Set 1: Method*

Sales volumes, margin of contribution, and shop traffic data are available for 173 of the 204 shops in the data set used Study 1. Because market factors such as local competition or local market volume can significantly influence sales volume, we used the conversion rates based on contribution margin as an indicator of sales performance.

To measure the strength of the association between salespeople attributes and sales performance (H1c and H3c), we first calculated simple bivariate correlations between the average attribute assessments of the two groups of assessors and the conversion rate of the shop in which the mystery visits and customer surveys took place. In order to test H1c and H3c, we then performed ordinary least squares (OLS) regression to measure the impact of the assessments of mystery shoppers and real customers on sales performance. As predictors, we used, as before, competence and interpersonal skill factors extracted from the five subjective salespeople attributes and the four objective scale items of salespeople behavior, all averaged at the shop level.

### *Data Set 1: Results*

The findings of our correlation analysis corroborate the assumed relationship between customer evaluations of salespeople and sales outcomes (H1b and H1c). Among the customer measures, the handout at exit item and the comprehensiveness of information show the highest correlations with the shop conversion rate (Pearson's r = .26 and r = .22, $p < .01$). The level of association between conversion rate and overall customer satisfaction with the salesperson is slightly lower (r = .18, $p < .01$). By contrast, we do not find evidence that MS assessments are associated with sales success, thus rejecting H3b and H3c. All correlations between conversion rate and the average MS assessments are non-significant and lower than .14.[1]

As a further test of H3c, we first estimated an OLS regression model in which we only included MS assessments as predictors, all averaged at the shop level and standardized (see Table 3). The estimated model is not significant ($F(6, 166) = 1.66$, $p > .05$), and the model fit is very poor ($R^2$ adj. = .02). None of the six predictors based on MS assessments are significantly associated with our measure of sales performance (logit of conversion rate), with one exception, the item handout at exit ($p < .05$). Given these results, we reject the assumption of a positive correlation between MS assessments of salespeople attributes and the salespeople's sales performance; i.e., H3c.

Second, we regressed logit-transformed conversion rates on customer assessments. Again, we used the competence and interpersonal skill factors and the four salespeople behavior items as predictor variables, all averaged at the shop level and then standardized. The estimated model is significant ($F(6, 166) = 10.44$, $p < .001$), and the overall fit is moderate ($R^2$ adj. = .25), as is the case in many studies relating customer metrics to customer purchasing behavior (Keiningham et al. 2015). The highest sales impact among the six predictors comes from the com-

[1] Correlation matrices are available on request from the authors.

petence skill factor ($\beta = .25$, $p < .001$), the handout at exit item ($\beta = .26$, $p < .001$) and welcome greeting ($\beta = -.23$, $p < .001$). The regression coefficients for the other salespeople variables are not significant (see Table 3). Notably, customer perceptions of salespeople's competence but not salespeople's interpersonal skills have a significant effect on sales outcomes. This finding confirms recent research that provides evidence that when choosing among service providers, customer's value competence more than moral and warmth traits (Kirmani et al. 2017). The negative coefficient for welcome greeting could be due to endogeneity, because the item is associated with shop traffic ($r = -.21$, $p < .01$); that is, if a shop is crowded, salespeople will be serving customers and more easily lose sight of new customers entering the store. A similar effect may hold true for the item handout at exit because observations on this item are correlated with customer purchase ($r = .38$, $p < .01$); that is, salespeople often add additional sales material (handouts) to the bag when purchases are made.

Finally, we estimated a regression model in which we included both types of assessments, from customers and mystery shoppers, to explain variation in sales performance. The fit of this extended model is similar to the customer data model ($R^2$ adj. = .24), and the regression coefficients for the customer- and MS-based predictors are similar to the estimates in the two separate models. The handout at exit item, based on observations from real customers, still has a significant effect on sales performance ($p < .001$), and it neutralizes the impact of the corresponding variable based on observations from mystery shoppers, which is no longer significant. Thus, in contrast with the customer variables, none of the predictors based on MS assessments have a significant effect on sales performance. All in all, these results confirm the established correlation between customer evaluations and sales results (H1b and H1c), whereas the assumed relationship between MS assessments and sales is not supported (H3b and H3c).

We performed various checks to test the robustness of these estimates. Among other checks, we estimated our regression model using different types of shops. For example, we used only 96 small and medium-sized shops managed by a small number of salespeople, mostly two or three. We also checked scale transformations of the MS assessments, such as the top-2-box customer satisfaction score (De Haan, Verhoef, and Wiesel 2015). In addition, we used other scale items from the MS survey that are not contained in the customer survey as predictor variables. In all cases, we find that assessments of mystery shoppers cannot explain the variations in shop performance.

### *Data Set 2: Method*

A different company gave us access to 10,205 MS protocols collected from visits in 490 authorized dealer shops from 2011 to 2014. Again, all shops have a similar design, they sell the same products that are exclusively from the service company, and the regular prices for the services are the same. In general, a shop manager or the shop's owner handles the personal selling in the shop. MS studies were performed twice per year (in spring and autumn), each time using two shopping scenarios and four mystery shoppers per shop. As MS studies occurred twice per year, we calculated the average monthly sales for each half-year period in which the MS occurred and used it as a measure for sales performance. In total, 8,851 records with MS and sales data from 471 shops are available for the years 2012–2014. We excluded 72 shops that (according to sales experts from our industry partner) had begun sales only recently, had a different sales focus, or were large shops with more than one salesperson. From the remaining 399 shops, we have 6,658 MS observations and half-year sales from which we used 90% (5,992 randomly selected observations) to estimate our model and 10% for a hold-out validation sample (666 cases).

The MS surveys used in the six waves of research from 2012 to 2014 each contain more than 60 question items similar to those in our first data set because the MS studies were performed by the same agency. We selected all items that had been used throughout all waves of research and that are related to the service and personal selling provided by the salespeople. In particular, we used 21 subjective measures of salespeople's attributes and behaviors. Other items in the survey include shop-related information and observations about the salespeople's behavior during the sales encounter. We used the information contained in the observational scale items to create three new behavioral metrics that can influence the outcome of a sales encounter: the number of questions asked, the number of sales materials used, and whether the salesperson offered additional products. Moreover, the survey contains the mystery shoppers' overall evaluations of the salespeople and the sales encounter ("My overall impression of the salespeople/the sales encounter was very good/very bad") as well as information about the type and location of the shop and the period when the MS occurred.

Again, the ratings of the 21 subjective salespeople attributes are significantly correlated with each other and with the overall satisfaction with the salesperson (see Appendix). Task-related items are highly correlated with interpersonal attributes. It is clear that the mystery shoppers have difficulties in differentiating between the various salespeople attributes, possibly because the evaluation task is complex and tedious or because the shoppers try to provide consistent judgments. As a consequence, the item responses appear to be affected by a halo effect and tend to be uniform (Dagger et al. 2013; Wirtz 2000). To examine the underlying structure of the ratings, we again performed a principal axis factor analysis with varimax rotation and extracted three factors with eigenvalues greater than 1, explaining 60% of the overall variance (Table 4). The first factor primarily comprises the interpersonal skills of personal selling, and the second factor reflects a salesperson's basic competence and selling knowledge. The third factor is also task-related and is mainly associated with *knowing how to sell* items, specifically needs assessment, providing extensive information, and arguing product benefits. Following Rentz et al. (2002), we designate the first factor as interpersonal skills, the second as technical skills, and the third as salesmanship skills, though the three factors do not sharply delineate among the different scale items. However, to control for multicollinearity and to reduce the number of items to a manageable size, we used the three skill factors to represent the 21 subjective scale items for further analysis.

Table 4
Principal axis factor analysis of MS item ratings (data set 2).

| | Scale items | Factor 1 "Interpersonal"skills | Factor 2 "Technical"skills | Factor 3 "Salesmanship"skills |
|---|---|---|---|---|
| (1) | Likability | **0.76** | 0.35 | 0.07 |
| (2) | Friendliness | **0.76** | 0.21 | 0.19 |
| (3) | Good atmosphere | **0.61** | 0.38 | 0.37 |
| (4) | Took concerns seriously | **0.61** | 0.48 | 0.17 |
| (5) | Engagement | **0.63** | 0.45 | 0.35 |
| (6) | Enthusiastic (products) | **0.52** | 0.32 | 0.41 |
| (7) | Took enough time | 0.45 | 0.40 | 0.50 |
| (8) | Expertise | 0.37 | **0.69** | 0.32 |
| (9) | Well informed | 0.34 | **0.69** | 0.36 |
| (10) | Comprehensiveness | 0.47 | **0.63** | 0.20 |
| (11) | Structured explanations | 0.39 | **0.62** | 0.32 |
| (12) | Responsiveness | **0.54** | 0.53 | 0.30 |
| (13) | Individual requirements | 0.48 | 0.47 | 0.45 |
| (14) | Offers best fit solution | 0.43 | **0.55** | 0.38 |
| (15) | Active talk | 0.47 | 0.50 | 0.45 |
| (16) | Needs assessment[d] | 0.06 | 0.09 | **0.70** |
| (17) | Product presentation | 0.14 | 0.20 | **0.54** |
| (18) | Extensive information provision | 0.34 | 0.48 | **0.64** |
| (19) | Benefits argumentation | 0.38 | 0.37 | **0.59** |
| 20) | Competitive advantages[d] | 0.12 | 0.15 | 0.48 |
| (21) | Countering objections[d] | 0.26 | 0.25 | 0.29 |
| | Proportion of variance | .22 | .20 | .17 |

n = 6,658. [d] Binary variable. Other items measured with reversed five-point Likert-type scale (1 = "strongly disagree," 5 = "strongly agree").
Bold values are the highest factor loadings of a scale item with a value greater than .5.

The number of MS observations per shop ranges from four to 25 (M: 16.69, SD: 7.16). Since our data describe a multilevel setting, in which the number of observational units (shops) is large and the amount of information per observational unit is limited, we use hierarchical Bayes regression (Allenby, Rossi, and McCulloch 2005; Kruschke 2015). In our model, we regress salespeople attributes on sales performance, as follows:

$$Y_i = X_i ß_i + \varepsilon_i,\ \varepsilon_i \sim iid\ N(0,\ \sigma_i NI_i), \quad i = 1, \ldots, m,$$

where $Y_i$ designates the sales performance (average monthly sales volume in a half-year period) of salesperson (shop) i, X is a vector of salesperson attributes and control variables with corresponding regression coefficients ß[2], and $\varepsilon$ is the individual error variance. The predictors contained in the vector X are the overall satisfaction with the salesperson, the three skill factors extracted from factor analysis, and the four measures of salespeople's behavior described previously. In addition, we included five dummy variables to control for the type of shop (specialized dealer or large sales area) and the size of the town where the shop is located (large, medium, or small).

[2] The regression equations for the m salespeople (shops) are tied together by the assumption of a common prior distribution for $ß_i$, with $ß_i \tilde{N}(\Delta, V_ß)$, where $\Delta$ is the matrix of estimated coefficients and $V_ß$ is the random-effects covariance matrix. We used the Gibbs sampler implemented in the bayesm package (Rossi, Allenby, and McCulloch 2005) for R (R Core Team 2015) to approximate the posterior distribution of the coefficients in our hierarchical model.

*Data Set 2: Results*

To examine whether MS assessments of salespeople are related to salespeople's performance (H3b and H3c), we first calculated bivariate correlations between sales volume and the salespeople attributes used in our model. The correlations are all low or even insignificant: Pearson's r varies between –.03 (df = 6.656, $p < .05$) for the interpersonal skills factor and .01 ($p > .05$) for whether the salesperson offered an additional product. The correlation between sales performance and the overall satisfaction with the salesperson is –.02 ($p > .05$). These results contradict our assumptions.

Second, we regressed sales performance on the overall satisfaction of the mystery shopper with the salesperson to test our assumption in H3b and then added mystery shoppers' assessments of salespeople attributes to test the assumption in H3c. In Table 5 (from the left-hand side to right-hand side), we report the posterior means of the estimated parameter distributions of (1) the estimated effect of mystery shoppers' overall satisfaction with a salesperson on sales performance, (2) the estimated performance effect of salespeople attributes, and (3) the parameter estimates if we use both mystery shoppers' overall satisfaction and salespeople attributes as predictors.

As Table 5 shows, mystery shoppers' overall satisfaction with the salesperson has, by itself, no significant effect on sales performance. The posterior mean estimate is zero (SD: .10). To check the model's predictive validity, we correlated the actual sales of the 666 shops in the hold-out sample with shop sales estimates calculated from the individual-level mean parameter estimates of the respective shops. Pearson's r between actual and

Table 5
Estimates of the performance model parameters (MS data, data set 2).

| Dependent variable: average monthly sales in half-year period | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Overall satisfaction with salesperson (H3b) | | | Salespeople attributes (H3c) | | | Overall satisfaction and salespeople attributes | | |
| Predictor (MS data) | Mean | 2.5% quantile | 97.5% quantile | Mean | 2.5% quantile | 97.5% quantile | Mean | 2.5% quantile | 97.5% quantile |
| Intercept | 41.67 | 37.28 | 46.76 | 41.11 | 36.33 | 46.36 | 40.55 | 35.93 | 45.87 |
| Overall satisfaction | .00 | −.20 | .21 | | | | .09 | −.30 | .49 |
| Interpersonal skills | | | | .04 | −.23 | .30 | −.01 | −.34 | .32 |
| Technical skills | | | | −.11 | −.38 | .15 | −.15 | −.46 | .15 |
| Salesmanship skills | | | | −.03 | −39 | .32 | −.06 | −.45 | .32 |
| Closing question[d] | | | | .03 | −.28 | .33 | .02 | −.30 | .33 |
| Additional offer[d] | | | | .28 | .02 | .55 | .29 | .02 | .56 |
| No. questions asked | | | | −.23 | −.57 | .11 | −.23 | −.58 | .12 |
| No. sales materials | | | | .01 | −.26 | .28 | .02 | −.26 | .30 |
| Specialized dealer[d] | 1.14 | .08 | 2.21 | 1.12 | .09 | 2.17 | 1.12 | .09 | 2.17 |
| Large sales area[d] | 3.88 | −.90 | 9.19 | 3.93 | −.57 | 8.50 | 4.11 | −1.12 | 9.10 |
| Town size L[d] | 9.33 | 2.85 | 15.71 | 9.61 | 2.78 | 16.35 | 10.12 | 3.60 | 16.28 |
| Town size M[d] | 5.71 | −.10 | 10.84 | 6.25 | −.12 | 11.89 | 6.76 | .62 | 12.14 |
| Town size S[d] | 5.32 | −.76 | 11.10 | 5.48 | −.79 | 11.23 | 6.31 | .08 | 12.68 |
| Pearson's r (hold-out) | .90 | | | .90 | | | .90 | | |

n = 5,992. [d] Binary variable. All predictor variables, except binary variables, were standardized by dividing by 2 standard deviations.
Posterior distribution approximated with 500,000 draws from which we retained every 10th draw for analysis. First 5,000 draws used for burn-in.
Effective sample sizes of salespeople variables range from 7,500 to 15,000.

estimates sales is .90. However, closer inspection reveals that the high correlation is exclusively due to the model's intercept term (the shop effect) and the control variables, whereas the overall satisfaction with the salesperson does not provide any incremental improvement in the model's fit. Thus, the variation in sales performance can be attributed to the specifics of the shops, not to mystery shoppers' satisfaction with the salesperson.

We find similar results when we use the mystery shoppers' assessments of salespeople attributes as predictors in the performance model (middle part of Table 5). The coefficients of the overall impression with a salesperson and the three skill factors are close to zero, as are the coefficients for asking a closing question and the number of sales materials used. Of all the predictor variables based on the assessments of mystery shoppers, we observe only one (small) effect on sales volume that is credibly different from zero: average sales volume is slightly higher in shops where mystery shoppers notice that salespeople offer an additional product. With regard to the control variables, we note that average sales volume is higher in specialized dealer shops and in shops with a large sales area and that sales are significantly lower in shops located in small cities (reference category).

When we correlate the actual sales of the 666 shops in the hold-out sample with sales estimates calculated from the mean parameter estimates of the respective shops, Pearson's r again equals .90. As before, this correlation comes primarily from the intercept and, to a lesser extent, from the control variables, not from the salespeople's variables. Apparently, the variation in shop sales can be predicted by the specifics of the shops and, to a lesser extent, by the type and location of the shop, whereas the salespeople attributes are neither directly related to sales performance nor mediated by the overall satisfaction with the salesperson. We obtain the same results when we use both the mystery shoppers' overall satisfaction and their assessments of the salespeople attributes as predictors of sales performance (right-hand side of Table 5) or when we employ maximum likelihood to estimate our multilevel model (Pinheiro et al. 2014). The parameter estimates are similar to the mean estimates in Table 5, and, among all salespeople items, offering an additional product is the only predictor that shows a significant effect on sales performance ($p < .05$).

### *Robustness Checks*

We performed additional analyses and robustness checks to explore the missing correlation. In particular, we examined several factors that could possibly moderate the effects of the salespeople attributes on sales performance, specifically the number of visits performed by a mystery shopper, the number of observations per shop, the product category brought up by mystery shoppers in the sales talk (sales scenario), and the region and size of the towns in which the shop is located. In all cases, we find that none of these factors interact with the overall satisfaction with the salesperson or the salespeople attributes.

We also used different performance measures as our dependent variable—for example, overall sales (business-to-consumer and business-to-business) and sales to new customers and the customer base, respectively—instead of total sales, but without success. In addition, we averaged the mystery shopper ratings for each observation period and used average values on the independent variables. Finally, we estimated our model with different sets of independent variables. In all cases, the results confirmed our findings that MS assessments are not credibly related to sales performance.

Table 6
Estimates of the performance model parameters (MS data, data set 3).

| Dependent variable: Logit-transformed average monthly conversion rate in half-year period | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Overall satisfaction with salesperson (H3b) | | | Salespeople attributes (H3c) | | | Overall satisfaction and salespeople attributes | | |
| Predictor (MS data) | Mean | 2.5% quantile | 97.5% quantile | Mean | 2.5% quantile | 97.5% quantile | Mean | 2.5% quantile | 97.5% quantile |
| Intercept | −3.36 | −.4.31 | −2.41 | −3.35 | −4.37 | −2.34 | −3.36 | −4.39 | −2.34 |
| Overall satisfaction | −.02 | −.10 | .06 | | | | −.01 | −.17 | .16 |
| Interpersonal skills | | | | −.04 | −.15 | .07 | −.03 | −.16 | .09 |
| Sales orientation skills | | | | .01 | −.10 | .11 | .01 | −.11 | .13 |
| Individual consultation | | | | .00 | −.11 | .11 | .01 | −.11 | .12 |
| Adaptive selling skills | | | | .04 | −.08 | .15 | .04 | −.09 | .18 |
| No. questions asked | | | | −.04 | −.15 | .07 | −.04 | −.16 | .08 |
| No. product benefits | | | | .03 | −.08 | .14 | .03 | −.09 | .14 |
| No. product advantages | | | | −.11 | −.21 | −.01 | −.11 | −.22 | −.01 |
| No. counterarguments | | | | .02 | −.08 | .12 | .02 | −.08 | .12 |
| Purchase power | −.04 | −.70 | .63 | −.05 | −.76 | .67 | −.05 | −.78 | .67 |
| Region 1[d] | .19 | −.52 | .91 | .22 | −.55 | .99 | .23 | −.55 | 1.01 |
| Region 2[d] | .13 | −.93 | 1.08 | .13 | −.80 | 1.14 | .13 | −.90 | 1.17 |
| Town size L[d] | .05 | −1.12 | 1.21 | .06 | −1.16 | 1.27 | .06 | −1.16 | 1.28 |
| Town size M[d] | .06 | −.82 | .91 | .03 | −.89 | .95 | .03 | −.90 | .96 |
| Shop traffic high[d] | −.18 | −.99 | .63 | −.19 | −1.06 | .67 | −.18 | −1.07 | .69 |
| Shop traffic medium[d] | −.19 | −1.01 | .62 | −.19 | −1.07 | .68 | −.19 | −1.07 | .69 |
| Pearson's r (hold-out) | .73 | | | .68 | | | .66 | | |

n = 1,328. [d] Binary variable. All predictor variables, except binary variables, were standardized by dividing by 2 standard deviations.
Posterior distribution approximated with 500,000 draws from which we retained every 10th draw for analysis. First 5,000 draws used for burn-in.
Effective sample sizes of salespeople variables range from 22,500 to 45,000.

### *Data Set 3: Method*

We replicated our analysis regarding hypotheses H3b and H3c with more detailed MS data from a competitor company. We gained access to sales data and 1,476 MS observations collected in the four quarters of 2016 and the first three quarters of 2017 for 130 shops operated by authorized dealers of the company. Again, the shops have a similar design, they sell the same products for the same prices, and, in most cases, one person serves the customers. Each shop was visited quarterly by two mystery shoppers independently, each of them using one of two predefined sales scenarios (mean number of MS observations per shop: 11.35, SD: 3.61).

Of the more than 70 question items in the MS surveys, some items provide meaningful information not covered in the other two data sets, such as the trustworthiness and adaptive behavior of the salesperson. Again, the surveys contain detailed behavioral observations of the mystery shoppers that we used to create four behavioral metrics: the number of questions asked, the number of product advantages, the number of benefits mentioned, and the number of counterarguments used by the salesperson. Moreover, the surveys include an overall evaluation of the salesperson ("The salesperson provided a very good consultation service"). As in Study 1, we subjected 22 highly correlated subjective salespeople measures to a principal axis factor analysis with varimax rotation and found four skill factors with eigenvalues greater than 1 that explain 80% of the variance: one factor representing interpersonal skills, one factor representing the salesperson's sales orientation, and two factors capturing more specific sales skills (i.e., individual, persuasive consultation and adaptive selling skills) (Franke and Park 2006; Spiro and Weitz 1990).

To measure sales performance, our partner company provided us with monthly data on sales volume and shop traffic. From this information, we calculated shop conversion rates (M: .033, SD: .01). We used the logit-transformed average monthly conversion rate of the shop in which the mystery visit took place as our measure of sales performance. Our predictors were the overall evaluation of the salesperson's consultation service, the four skill factors, and the four observational metrics (e.g., the number of questions asked). In addition, we used several control variables, including the location of the shop and the purchasing power of the town in which the shop is located.

### *Data Set 3: Results*

We randomly assigned 90% of the cases (1,328 observations) to model estimation and 10% of the cases to a hold-out validation sample. The mean effect sizes of the overall satisfaction with a salesperson and all salespeople attributes are again close to zero (see Table 6). The number of product advantages mentioned by the salesperson has the highest mean estimate (–.11) among the salespeople variables in absolute terms. This implies a reduction of the average monthly conversion rate; that is, the more product advantages a salesperson needs to bring up in a sales encounter, the less likely he or she is to close the deal.

When we correlate the logit-transformed conversion rates of the shops in the holdout sample with estimates based on the shop-level means of the parameter estimates, the marginal contribution of the salespeople variables to this correlation is close

to zero. As in the second data set, the predictive quality of the model comes from the intercept and the control variables, not from the salespeople predictors. This finding does not change when we include more predictor variables (e.g., the type of product offered), use quadratic terms of the salespeople attributes, or change the specification of the performance measure. Employing a mixed-effects model and maximum likelihood to estimate the model parameters also does not substantially change the results, because we do not find any salespeople predictors to be significant at the 5% level, except one (i.e., the number of benefits mentioned by the salesperson). The effect of this variable is lower than some of the control variables, and it is negative, which contradicts conventional wisdom. Thus, contrary to our expectations, even the availability of extensive MS information is not sufficient to establish a meaningful relationship between MS assessments and objective sales performance.

We conclude from the results of our checks that our findings are robust. In summary, MS data, collected by professional MS organizations, are not related to sales performance (with the exception of a weak effect of cross-selling). Our assumption was that mystery shoppers' satisfaction with salespeople and their assessments of the salespeople attributes would be related to a salesperson's sales performance. Our data empirically contradict this assumption, and therefore we reject our basic hypotheses (H3b and H3c).

## Complementary Studies

The missing relationship between MS assessments and corresponding evaluations of real customers explains to a large extent why MS data are not predictive of sales performance. In addition, we examined other possible root causes, including the reliability of the MS assessments and whether salespeople are able to unmask mystery shoppers.

### *Reliability*

To assess the interrater reliability of the MS assessments, we calculated intraclass correlations (ICC) for the second and third data set. In the second data set, the ICC(1), which measures the reliability of individual rating values, is low for the 18 subjective salespeople attributes measured with Likert-type scales (see Table 4). Calculated with a one-way analysis of variance (Bliese 2000), the mean ICC(1) for the 18 scale items across all observations is .03 (SD: .01), with individual values ranging from .01 (enthusiastic) to .04 (well informed). All these correlations are far below the threshold for good ICC values of .6 (Hallgren 2012).

The ICC(2), which estimates the reliability of the mean shop ratings and is a function of the ICC(1) and group size, has an average value of .32 (SD: .06). Individual values for the 18 subjective items range from .18 to .41. The average number of 16.69 mystery shopper ratings per shop is not sufficient to raise the reliability of the mean ratings to the critical value of .6. When we calculate the ICC for smaller periods of observation (i.e., years or half-years), we find that the ICC(1) values are approximately at the same level as in the total sample while the ICC(2) values, due to the lower number of observations per shop, decrease.

The ICC values of the MS ratings in our third data set are higher than those in the second data set, though they are still low when measured against the benchmark level for good values. The mean ICC(1) value that we calculated for 24 subjective salespeople measures is .08 (SD: .04), with individual values ranging from .01 (language easy to understand) to .13 (tried hard to convince me). The ICC(2) values range between .10 and .63 (M: .48, SD: .16).

Standard ICC coefficients tend to underestimate interrater reliability in ill-structured measurement settings that are not from fully crossed or nested experimental designs (Putka et al. 2008). Therefore, we used another measure, the G-coefficient proposed by Putka et al. (2008), to cross-check our results. Because calculation of the G-coefficient requires information about the identity of the raters, we used our second data set, in which we have the anonymized identity codes of the 120 mystery shoppers who participated in the MS studies. For the 18 subjective measures in our second data set, the G-coefficient ranges from .13 to .35 (M: .21, SD: .07). Though higher than the ICC(1), these values confirm the low interrater reliability of the individual mystery shopper assessments.

As a complement, we also calculated a measure of intrarater reliability using ratings by 43 mystery shoppers from our second data set who evaluated the same shop twice, usually after one year. We calculated the correlation coefficients from 931 duplicated evaluations using the ratings on the 18 subjective salespeople items. The overall mean of the Pearson's correlation coefficients is .81 (SD: .11), with 25% of the correlations ranging between .89 and 1. Thus, while repeated assessments of the same mystery shopper show a high degree of consistency, consensus in the assessments of different mystery shoppers, which is most relevant for the reliability of MS data, is low.

### *Check of Whether Mystery Shoppers Are Uncovered*

To check whether salespeople are able to uncover mystery shoppers, a research assistant interviewed 19 salespeople in the stores of the partner company that provided the third data set. All stores had been visited by three mystery shoppers within six weeks before. When asked how likely mystery shoppers can be uncovered, eight of the 17 respondents noted that mystery shoppers can be unmasked with a probability of higher than 50%, and more than half the respondents claimed that they had already detected mystery shoppers in the past. Asked when the last time was that they had observed mystery shoppers in their store, only four salespeople declared that they had been visited by a mystery shopper in the last three months. Finally, when we asked the four salespeople to provide a rough description of the mystery shoppers they had uncovered, we received only two responses, one of which was traceable and matched the profile of one of the mystery shoppers. Given these results, we have no reason to believe that unmasking mystery shoppers is a root cause for the missing correlation between MS data and sales performance.

## Conclusions, Limitations, and Avenues for Further Research

Our research reveals that the level of agreement between the assessments of mystery shoppers and those of real customers is low and that MS assessments are not effective in predicting customer satisfaction. In addition to a low level of interrater reliability, we consider this finding a major root cause for the missing relationship between MS assessments of a salesperson and the salesperson's sales performance. While we confirm the relationship between customer evaluations and sales performance, our results indicate that companies do not benefit (in the sense of learning about additional drivers of sales performance based on MS) from conducting MS in parallel to customer satisfaction studies.

Although our findings are based on large-scale data from three different companies, our study has several limitations. First, our analysis comprised consultative retail settings in which the subjective aspects of personal selling play a major role for customer satisfaction and sales results. Our results cannot be extrapolated to other sectors in which the human factor is not a major determinant of sales performance. Second, because all three partner companies offer intensive sales training to their retail outlets, the differences in the quality of the personal selling within the respective retail networks may not be pronounced enough to explain the variation in sales performance. Benchmarking shops from different companies may yield more heterogeneous MS assessments and perhaps different results. Another limitation is our measure of sales performance. Although most of the shops in our second data set are similar in size, differences in the number of shop visitors may significantly influence sales volume. Conversion rate, our dependent variable in the two other data sets, is a more accurate performance measure but nevertheless may be biased by, for example, a high number of shop visitors who have service requests but do not want to buy.

For privacy reasons, our data do not contain information about other characteristics of the assessed salespeople and the mystery shoppers themselves. Those characteristics could help better understand the interaction between a salesperson and mystery shopper. Finally, most of the shops in two of our data sets were usually managed by only one person, the owner or the shop manager. However, our MS studies do not control for cases in which shop managers were absent and mystery shoppers interviewed other salespeople, temporary workers, or family members.

Regardless of these limitations, our study underscores the need to improve the MS method, and it provides strong paths for further research. We recommend an agenda for practitioners and researchers that particularly addresses the following topics:

1 *Reliability of MS data:* Mystery research agencies claim that intensive training is at the core of their business, and the agencies that conduct mystery research for our three partner companies are highly recognized in the industry. However, from our data, the question arises whether cognitive overload or a lack of diligence or competence on the part of mystery shoppers is seriously undermining the reliability of MS results. Stringent quality controls are necessary to guarantee reliable data. Our results indicate the importance of cross-validating assessments of mystery shoppers with information from other sources (e.g., customers, sales experts, sales trainers, self-reports of salespeople) and with objective data on service quality and sales performance. Cross-checking MS results with observations from other sources could also increase the acceptance of MS studies among salespeople, and it would put pressure on the research agencies and mystery shoppers to deliver reliable data.

2 *Scale development:* In line with Keiningham et al. (2015), who advocate the use of relative metrics in customer satisfaction surveys, it would be useful to determine whether relative scales improve predictive accuracy with respect to sales performance. Relative measures could produce more mixed evaluations (Judd et al. 2005) and also help reduce the risk of single-sided ratings and ceiling effects.

3 *Objective measurement of personal selling:* The key advantage of the MS method over customer surveys is that mystery shoppers are able to notice detailed aspects of the personal selling provided by salespeople that real customers usually cannot recall. MS studies should focus on this advantage instead of providing subjective evaluations of salespeople, which may not be consistent with evaluations of real customers. To establish a relationship between salespeople behavior and sales performance, MS could be carried out to measure if and how salespeople try to influence customers' buying behavior. To capture salespeople's competence and interpersonal skills as well as the influence tactics they use, mystery shoppers would need to observe their selling approach, including their main questions and responses and listening behavior (DeCormier and Jobber 1993; McFarland, Challagalla, and Shervani 2006). Developing such a scaling of personal selling requires more preparation work than conventional MS surveys, and processing such extensive information would be demanding. However, the mystery shoppers could shift their attention from subjective impression building to objective observation.

4 *Technology:* Our results lend support to Finn's (2001) assertion that store evaluations based only on three to four mystery visits are inadequate for reliable benchmarking. However, according to our own discussions with sales managers, companies are generally not poised to perform 20 or more visits per shop and wave of research. Therefore, other tools must be considered. One possible option is to record mystery visits with audio or video devices, subject to the express approval of all the parties involved and in accordance with the legal requirements. The recordings of the sales encounters could then be analyzed and evaluated—in addition to or instead of mystery shoppers—by sales experts, managers, and panels of real customers. As a result, the MS task would become less demanding, and the risk of information overload and biased evaluations would diminish considerably. Above all, the information provided by an MS visit could be analyzed by several people, from different perspectives and with different objectives. Cross-validation of evaluations, as rec-

ommended previously, could be easily implemented, as well as the complex measurement of influence tactics that might be too difficult for an average mystery shopper. Thus, the data efficiency of MS surveys would significantly benefit from sales talk recordings, as well as the accuracy of the results.

Another step to manage the unstructured data contained in recordings of MS visits is to use advanced techniques of text analysis. Sales talks can be translated into text, and text data from MS visits could be structured and analyzed with automatic text analysis software. Although the evaluation of salespeople attributes and behavior based on recordings of sales encounters is far more complex than a simple sentiment or a topic analysis, it is realistic to assume that appropriate software tools will be available in the foreseeable future. These tools will then be able to analyze the large amount of unstructured data contained in sales talk recordings efficiently, objectively, and quickly. These tools should enable the extraction from MS data of the pattern of influence of salespeople attributes and behavior on sales performance.

Research on the aforementioned topics will help determine whether MS can provide any meaningful contribution to the management of salespeople. To justify the high costs of the tool, it should be possible to relate MS assessments to sales performance and/or customer satisfaction. From our conversations with managers of the cooperating companies, this seems a precondition for the tool's acceptance among salespeople. Most of the salespeople we interviewed value the information provided by mystery research. If future research does not show that MS results are based on the traceable judgments of mystery shoppers, acceptance of the technique may dramatically suffer.

## Conflicts of interest

None.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Acknowledgement

We are grateful to René Algesheimer, Günter Hitsch, Arvind Rangaswamy, Klaus Miller, Jochen Reiner, as well as to the participants of the Doctoral Seminar of Goethe University Frankfurt for their helpful feedback on previous versions of the paper. Furthermore, we appreciate the many very helpful comments and suggestions of the editor and the anonymous reviewers. This research was supported by the University Research Priority Program "Social Networks" at the University of Zurich.

## Appendix. Correlation matrix.[a]

| | | (0) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) | (14) | (15) | (16) | (17) | (18) | (19) | (20) | (21) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (0) | Overall satisfaction | | | | | | | | | | | | | | | | | | | | | | |
| (1) | Likability | .64 | | | | | | | | | | | | | | | | | | | | | |
| (2) | Friendliness | .61 | .68 | | | | | | | | | | | | | | | | | | | | |
| (3) | Good atmosphere | .69 | .66 | .62 | | | | | | | | | | | | | | | | | | | |
| (4) | Took concerns seriously | .59 | .64 | .58 | .60 | | | | | | | | | | | | | | | | | | |
| (5) | Engagement | .67 | .65 | .63 | .66 | .69 | | | | | | | | | | | | | | | | | |
| (6) | Enthusiastic (products) | .56 | .51 | .58 | .57 | .55 | .63 | | | | | | | | | | | | | | | | |
| (7) | Took enough time | .64 | .51 | .52 | .60 | .58 | .65 | .56 | | | | | | | | | | | | | | | |
| (8) | Expertise | .64 | .56 | .51 | .60 | .60 | .66 | .57 | .59 | | | | | | | | | | | | | | |
| (9) | Well informed | .63 | .54 | .50 | .61 | .58 | .64 | .56 | .58 | .78 | | | | | | | | | | | | | |
| (10) | Comprehensiveness | .61 | .60 | .52 | .60 | .63 | .64 | .53 | .56 | .67 | .65 | | | | | | | | | | | | |
| (11) | Structured explanations | .63 | .54 | .48 | .60 | .58 | .62 | .51 | .59 | .65 | .66 | .66 | | | | | | | | | | | |
| (12) | Responsiveness | .63 | .61 | .57 | .63 | .66 | .68 | .54 | .60 | .64 | .64 | .64 | .63 | | | | | | | | | | |
| (13) | Individual requirements | .62 | .56 | .53 | .62 | .64 | .69 | .56 | .61 | .63 | .62 | .61 | .61 | .71 | | | | | | | | | |
| (14) | Offers best fit solution | .60 | .54 | .50 | .60 | .62 | .66 | .58 | .60 | .65 | .65 | .60 | .64 | .65 | .67 | | | | | | | | |
| (15) | Active talk | .64 | .56 | .53 | .63 | .61 | .71 | .59 | .64 | .66 | .65 | .61 | .63 | .67 | .68 | .68 | | | | | | | |
| (16) | Needs assessment[d] | .32 | .12 | .19 | .31 | .20 | .33 | .33 | .41 | .29 | .32 | .22 | .28 | .31 | .44 | .34 | .40 | | | | | | |
| (17) | Product presentation | .39 | .22 | .22 | .39 | .27 | .37 | .40 | .41 | .38 | .39 | .31 | .35 | .32 | .39 | .35 | .38 | .40 | .35 | | | | |
| (18) | Extensive information provision | .63 | .48 | .49 | .64 | .53 | .65 | .58 | .68 | .67 | .69 | .59 | .64 | .62 | .68 | .65 | .69 | .52 | .35 | | | | |
| (19) | Benefits argumentation | .62 | .45 | .50 | .60 | .49 | .60 | .62 | .65 | .60 | .62 | .52 | .58 | .54 | .57 | .61 | .63 | .46 | .38 | .71 | | | |
| (20) | Competitive advantages[d] | .33 | .20 | .23 | .34 | .22 | .28 | .26 | .34 | .29 | .33 | .25 | .30 | .30 | .34 | .31 | .34 | .35 | .39 | .40 | .37 | | |
| (21) | Countering objections[d] | .37 | .30 | .30 | .35 | .34 | .38 | .36 | .34 | .35 | .36 | .35 | .35 | .37 | .37 | .36 | .38 | .21 | .26 | .37 | .39 | .30 | |

[a]Pearson's correlation coefficients. All correlations are significant ($p < .01$).
[d]Binary scale item. All other items measured on a reversed five-point Likert-type scalee.

## References


Ahearne, Michael, John Mathieu and Adam Rapp (2005), "To Empower or Not to Empower Your Sales Force? An Empirical Examination of the Influence of Leadership Empowerment Behavior on Customer Satisfaction and Performance," *Journal of Applied Psychology,* 90 (5), 945–55.

Ailawadi, Kusum L., Scott A. Neslin, Y. Jackie Luan and Gail Ayala Taylor (2014), "Does Retailer CSR Enhance Behavioral Loyalty? A Case for Benefit Segmentation," *International Journal of Research in Marketing,* 31 (2), 156–67.

Allenby, Greg M., Peter E. Rossi and Robert E. McCulloch (2005) "Hierarchical Bayes Models: A Practitioners Guide" (accessed May 7, 2018), [available at SSRN: https://ssrn.com/abstract=655541 or https://doi.org/10.2139/ssrn.655541].

Anderson, Eugene W. and Vikas Mittal (2000), "Strengthening the Satisfaction-Profit Chain," *Journal of Service Research,* 3 (2), 107–20.

Babin, Laurie A., Barry J. Babin and James S. Boles (1999), "The Effects of Consumer Perceptions of the Salesperson, Product and Dealer on Purchase Intentions," *Journal of Retailing and Consumer Services,* 6 (2), 91–7.

Bliese, Paul D. (2000), "Within-Group Agreement, Non-Independence, and Reliability: Implications for Data Aggregation and Analysis," in *Multilevel Theory, Research, and Methods in Organizations*, Klein Katherine J. and Kozlowski Steve W. J., eds. San Francisco, CA: Jossey-Bass, Inc, 329–81.

Brady, Michael K. and Joseph J. Cronin (2001), "Customer Orientation: Effects on Customer Service Perceptions and Outcome Behaviors," *Journal of Service Research,* 3 (3), 241–51.

Brexendorf, Tim Oliver, Silke Mühlmeier, Torsten Tomczak and Martin Eisend (2010), "The Impact of Sales Encounters on Brand Loyalty," *Journal of Business Research,* 63 (11), 1148–55.

Calvert, Philip (2005), "It's a Mystery: Mystery Shopping in New Zealand's Public Libraries," *Library Review,* 54 (1), 24–35.

Carrillat, François A., Fernando Jaramillo and Jay Prakash Mulki (2009), "Examining the Impact of Service Quality: A Meta-Analysis of Empirical Evidence," *Journal of Marketing Theory and Practice,* 17 (2), 95–110.

Crosby, Lawrence A., Kenneth R. Evans and Deborah Cowles (1990), "Relationship Quality in Services Selling: An Interpersonal Influence Perspective," *Journal of Marketing,* 54 (7), 68–81.

Dagger, Tracey S., Peter J. Danaher, Jillian C. Sweeney and Janet R. McColl-Kennedy (2013), "Selective Halo Effects Arising from Improving the Interpersonal Skills of Frontline Employees," *Journal of Service Research,* 16 (4), 488–502.

De Haan, Evert, Peter C. Verhoef and Thorsten Wiesel (2015), "The Predictive Ability of Different Customer Feedback Metrics for Retention," *International Journal of Research in Marketing,* 32 (2), 195–206.

DeCormier, Ray A. and David Jobber (1993), "The Counselor Selling Method: Concepts and Constructs," *Journal of Personal Selling & Sales Management,* 13 (4), 39–59.

Erstad, Margaret (1998), "Mystery Shopping Programmes and Human Resource Management," *International Journal of Contemporary Hospitality Management,* 10 (1), 34–8.

ESOMAR (2005), *ESOMAR World Research Codes and Guidelines. Mystery Shopping Studies,* (accessed January 22, 2018), [available at https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR_Codes-and-Guidelines_MysteryShopping.pdf].

Finn, Adam (2001), "Mystery Shopper Benchmarking of Durable-Goods Chains and Stores," *Journal of Service Research,* 3 (4), 310–20.

Finn, Adam and Ujwal Kayandé (1999), "Unmasking a Phantom: A Psychometric Assessment of Mystery Shopping," *Journal of Retailing,* 75 (2), 195–217.

Franke, George R. and Jeong-Eun Park (2006), "Salesperson Adaptive Selling Behavior and Customer Orientation: A Meta-Analysis," *Journal of Marketing Research,* 43 (4), 693–702.

Gabler, Colin B., Jessica L. Ogilvie, Adam Rapp and Daniel G. Bachrach (2017), "Is There a Dark Side of Ambidexterity? Implications of Dueling Sales and Service Orientations," *Journal of Service Research,* 20 (4), 379–92.

Gelman, Andrew (2008), "Scaling Regression Inputs by Dividing by Two Standard Deviations," *Statistics in Medicine,* 27, 2865–73.

Gomez, Miguel I., Edward W. McLaughlin and Dick R. Wittink (2004), "Customer Satisfaction and Retail Sales Performance: An Empirical Investigation," *Journal of Retailing,* 80 (4), 265–78.

Grewal, Dhruv and Arun Sharma (1991), "The Effect of Salesforce Behavior on Customer Satisfaction: An Interactive Framework," *Journal of Personal Selling & Sales Management,* 11 (2), 13–23.

Grewal, Dhruv, Michael Levy and Greg W. Marshall (2002), "Personal Selling in Retail Settings: How Does the Internet and Related Technologies Enable and Limit Successful Selling?," *Journal of Marketing Management,* 18, 301–31.

Hallgren, Kevin A. (2012), "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial," *Tutorials in Quantitative Methods for Psychology,* 8 (1), 23–34.

Heskett, James L., Thomas O. Jones, Garry W. Loveman, W. Earl Sasser and Leonard A. Schlesinger (1994), "Putting the Service-Profit Chain to Work," *Harvard Business Review,* 72 (2), 164–74.

Hoekstra, Janny C., Annette Ammeraal and Peter S.H. Leeflang (2014), "Do Mystery Callers Really Represent Customers?," in *Conference paper, European Marketing Academy Conference*.

Homburg, Christian, Michael Müller and Martin Klarmann (2011), "When Should the Customer Really be King? On the Optimum Level of Salesperson Customer Orientation in Sales Encounters," *Journal of Marketing,* 75 (2), 55–74.

Hunneman, Auke, Peter C. Verhoef and Laurens M. Sloot (2015), "The Impact of Consumer Confidence on Store Satisfaction and Share of Wallet Formation," *Journal of Retailing,* 91 (3), 516–32.

Jaramillo, Fernando, François A. Carrillat and William B. Locander (2005), "A Meta-Analytic Comparison of Managerial Ratings and Self-Evaluations," *Journal of Personal Selling & Sales Management,* 25 (4), 315–28.

Judd, Charles M., Laurie James-Hawkins, Vincent Yzerbyt and Yoshihisa Kashima (2005), "Fundamental Dimensions of Social Judgment: Understanding the Relations Between Judgments of Competence and Warmth," *Journal of Personality and Social Psychology,* 89, 899–913.

Keiningham, Timothy L., Bruce Cooil, Edward C. Malthouse, Alexander Buoye, Lerzan Akzoy, Arne de Keiser and Bart Lavivière (2015), "Perceptions are Relative," *Journal of Service Management,* 26 (1), 2–43.

Kirmani, Amna, Rebecca W. Hamilton, Debora V. Thompson and Shannon Lantzy (2017), "Doing Well Versus Doing Good: The Differential Effect of Underdog Positioning on Moral and Competent Service Providers," *Journal of Marketing,* 81 (1), 103–17.

Kruschke, John K. (2015), *Doing Bayesian Analysis. A Tutorial with Jags, R, and Stan,* 2nd ed. Academic Press, Elsevier.

Leibowotz, Josh (2010), "Rediscovering the Art of Selling," *McKinsey Quarterly,* 2, 117–9.

Levy, Michael and Arun Sharma (1993), "Relationships Among Measures of Retail Salesperson Performance," *Journal of the Academy of Marketing Science,* 3, 231–8.

Lowndes, Michelle and John Dawes (2001), "Do Distinct SERVQUAL Dimensions Emerge from Mystery Shopping Data? A Test of Convergent Validity," *Canadian Journal of Program Evaluation,* 16 (2), 41–54.

Macintosh, Gerard and Lawrence S. Lockshin (1997), "Retail Relationships and Store Loyalty: a Multi-Level perspective," *International Journal of Research in Marketing,* 14 (5), 487–97.

Mattson, Jan (2011), "Strategic Insights from Mystery Shopping in B2B Relationships," *Journal of Strategic Marketing,* 20 (4), 313–22.

McFarland, Richard G., Goutam N. Challagalla and Tasadduq A. Shervani (2006), "Influence Tactics for Effective Adaptive Selling," *Journal of Marketing,* 70 (4), 103–17.

Morrison, Lisa J., Andrew M. Colman and Carolyn C. Preston (1997), "Mystery Customer Research: Cognitive Processes Affecting Accuracy," *Journal of the Market Research Society,* 39 (2), 349–61.

MSPA (2018), *Mystery Shopping – How Big is The Market,* (accessed June 6, 2018), [available at https://www.mspa-ea.org/news/newsitem/58-mystery-shopping-how-big-is-the-market.html].

——— (2018), *General Mystery Shopping Industry Information,* MSPA Europe, Mystery Shopping Providers Association (accessed January 22, 2018), [available at http://www.mspa-eu.org/contact.html]

Neff, Jack (2008), "Pick a Product: 40% of Public Decide in Store," *Advertising Age,* 79 (29) (accessed May 7, 2018), [available at https://adage.com/article/news/pick-a-product-40-public-decide-store/129924]

Peterman, Karen and Denise Young (2015), "Mystery Shopping: An Innovative Method for Observing Interactions with Scientists during Public Science Events," *Visitor Studies,* 18 (1), 83–102.

Pinheiro, José, Douglas Bates, Sikat DebRoy and Deepayan Sarkar (2014), *R Core Team (2014) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-117,* (accessed May 7, 2018), [available at http://CRAN.R-project.org/package=nlme (2014)].

Plouffe, Christopher R., Willy Bolander and Joseph A. Cote (2014), "Which Influence Tactics Lead to Sales Performance? It Is a Matter of Style," *Journal of Personal Selling & Sales Management,* 34 (2), 141–59.

Price, Linda L., Eric J. Arnould and Sheila L. Deibler (1995), "Consumers' Emotional Responses to Service Encounters: The Influence of the Service Provider," *International Journal of Service Industry Management,* 6 (3), 34–63.

Putka, Dan J., Huy Le, Rodney A. McCloy and Tirso Diaz (2008), "Ill-structured Measurement Designs in Organizational Research: Implications for Estimating Interrater Reliability," *Journal of Applied Psychology,* 93 (5), 959–81.

R Core Team (2015), *R: A Language and Environment for Statistical Computing,* Vienna, Austria: R Foundation for Statistical Computing (accessed January 22, 2018), [available at https://www.R-project.org/]

Rentz, Joseph O., David Shepherd, Armen Tashchian, Pratibha A. Dabholkar and Robert T. Ladd (2002), "A Measure of Selling Skill: Scale Development and Validation," *Journal of Personal Selling & Sales Management,* 22 (1), 13–21.

Rossi, Peter E., Greg M. Allenby and Rob McCulloch (2005), *Bayesian Statistics and Marketing,* New York: Wiley.

Snijders, Tom and Roel Bosker (2012), *Multilevel Analysis: An Introduction to Basic and Applied Multilevel Analysis,* 2nd ed. Thousand Oaks, CA: Sage.

Spiro, Rosann L. and Barton A. Weitz (1990), "Adaptive Selling: Conceptualization, Measurement, and Nomological Validity," *Journal of Marketing Research,* 27 (2), 61–9.

Swan, John E. and Richard L. Oliver (1991), "An Applied Analysis of Buyer Equity Perceptions and Satisfaction with Automobile Salespeople," *Journal of Personal Selling & Sales Management,* 11 (2), 15–26.

Sweeney, Jillian C., Geoffrey N. Soutar and Lester W. Johnson (1997), "Retail Service Quality and Perceived Value: A Comparison of Two Models," *Journal of Retailing and Consumer Services,* 4 (1), 39–48.

Van der Wiele, Ton, Martin Hesselink and Jos Van Iwaarden (2005), "Mystery Shopping: A Tool to Develop Insight into Customer Service Provision," *Total Quality Management & Business Excellence,* 16 (4), 529–41.

Van Dolen, Willemijn, Jos Lemmink, Ko de Ruyter and Ad de Jong (2002), "Customer-Sales Employee Encounters: A Dyadic Perspective," *Journal of Retailing,* 78 (4), 265–79.

Verbeke, Willem J., Bart Dietz and Ernst Verwaal (2011), "Drivers of Sales Performance: A Contemporary Meta-Analysis. Have Salespeople Become Knowledge Brokers?," *Journal of the Academy of Marketing Science,* 39 (3), 407–28.

Westbrook, Robert A. (1981), "Sources of Consumer Satisfaction with Retail Outlets," *Journal of Retailing,* 57 (3), 68–85.

Wilson, Alan M. (1998a), "The Uses of Mystery Shopping in the Measurement of Service Delivery," *Service Industries Journal,* 18 (3), 148–63.

——————— (1998b), "The Role of Mystery Shopping in the Measurement of Service Performance," *Managing Service Quality,* 8 (6), 414–20.

——————— (2001), "Mystery Shopping: Using Deception to Measure Service Performance," *Psychology & Marketing,* 18 (7), 721–34.

Wilson, Alan M. and Justin Gutmann (1998), "Public Transport: The Role of Mystery Shopping in Public Transport Decisions," *Journal of the Market Research Society,* 40 (4), 285–93.

Wirtz, Jochen (2000), "An Examination of the Presence, Magnitude and Impact of Halo on Consumer Satisfaction Measures," *Journal of Retailing and Consumer Services,* 7, 89–99.

Xu, Lifang and Shijie He (2014), "Analysis on The Survey Method of Mystery Shopping in Hospitality Management," E-Commerce," *E-Business ad E-Service,* 1, 221–5.