The Reliability of Mystery Shopper Reports: The Effects of Disconfirmed Expectancies and Exposure to Misinformation

Faculty of Behavioral Sciences, University of Twente, Enschede, the Netherlands

Student	**S.J. (Sijbrand) van der Tang**
Studentnumber	*s*0148016
Supervisor (First)	**Dr. J.J. (Joris) van Hoof**
University of Twente	
Supervisor (Second)	**Dr. J.F. (Jordy) Gosselt**
University of Twente	

Keywords: Mystery Shopper Reports; Reliability; Disconfirmed Expectancies; Exposure to Misinformation; Speak up

Number of words: abstract: 345 | main text: 14,028

OBJECTIVES: An important obstacle impeding the reliability of mystery shopper reports is researcher cognition bias, as mystery shoppers are the research instrument. This study investigates to what extend mystery shopping reports are reliable by investigating effects of two forms of researcher cognition bias: *disconfirmed expectancies* and *exposure to misinformation*.

METHOD: Using the method of mystery shopping, 63 mystery shoppers were divided over four conditions in a 2 * 2 field experiment (no disconfirmed expectancies vs. disconfirmed expectancies and no misinformation vs. misinformation). Instructed with a (manipulated) checklist, mystery shoppers were instructed to remember and report exact prices of five products from a local supermarket. Also, nineteen mystery shoppers participated in follow up interviews. Using an evolutionary interview structure, the goal of the interview was to offer explanation and meaning to findings of mystery shop visits. The follow up interviews proved a useful method for exposing experienced difficulties, such as forgotten or misremembered products.

RESULTS: Out of 315 products observed, 217 times mystery shoppers reported correct, which represents an average of 3.33 out of five correct reports per mystery shopper. Mystery shoppers who were disconfirmed in their expectancies showed a halo-effect which negatively influenced correct reporting rates of surrounding items on checklists. The effects of exposure to misinformation were limited to manipulated products and did not show forms of a halo-effect. Mystery shoppers who were not confronted with a researcher cognition bias reported a significant higher number of correct reports (4.24). One third of mystery shoppers spoke up to the research leader about experienced difficulties. Follow up interviews showed that a lack of self-confidence in own observations and a lack of intrinsic motivation are reasons for mystery shoppers to deliberately withhold information about experienced difficulties.

CONCLUSIONS: Results suggest that effects of researcher cognition bias have a significant negative effect on the reliability of mystery shopper reports. Inaccuracy of checklists resulted in a rise of incorrect reports, also for existing products. Moreover, two third of mystery shoppers did not speak up about experienced difficulties. Deliberately withholding information about experienced difficulties impedes the reliability of mystery shopper reports.

INTRODUCTION

Mystery shopping is the use of participant observers to monitor and report on a service experience (Wilson, 1998). The Mystery Shoppers Providers Association (MSPA) defines mystery shopping as “the practice of using trained shoppers to anonymously evaluate customer service, operations, employee integrity, merchandising, and product quality”. Mystery shopping is a 1.5 billion dollar industry worldwide, with a current estimate that there are 1.5 million mystery shoppers (MSPA, 2013). The scope of mystery shopping spans many industries, such as top retail outlets, restaurants, banks, theatres, travel companies, hotels, airlines and leisure organizations for the purpose of the measurement of service performance (Dawes & Sharp, 2000; Finn & Kayandé, 1999; van der Wiele, Hesselink, & van Iwaarden, 2005; Wilson, 1998; Wilson, 2001). For example, in Gosselt, van Hoof, de Jong and Prinsen (2007) under aged adolescent mystery shoppers tried to buy alcoholic beverages in supermarkets and liquor stores to measure levels of compliance among sales personnel. Mystery shopping was also used to evaluate use of simulated patients in assessment of clinical safety for telephone primary care service (Moriarty, McLeod & Dowell, 2003). Ford, Latham and Lennox (2011) used mystery shoppers to create a new tool for coaching employee performance improvement. Due to wide variance in use of mystery shopping, implications of mystery shopping reports also varies.

The implications of mystery shopping reports can vary from appraisal of staff with high performance (Wilson, 2001), discipline staff with low performance, appraisal or warning for a certain branch, to ultimately the dismissal of staff or a business. In the study of Gosselt et al. (2007), mystery shop reports eventually resulted in exacerbation of sales protocol for alcoholic beverages in the Netherlands. Mystery shopping reports resulted in regular use of simulated patients to evaluate limitations of health services in New Zealand (Moriarty, McLeod & Dowell, 2003). In the study of Ford, Latham and Lennox (2011) mystery shopping reports resulted in regular use of mystery shoppers to assess employee performance on predefined and measurable performance factors in organizational settings.

Mystery shopping has many advantages compared to other methods of service evaluation, such as customer evaluation. An important advantage of mystery shopping is the quality of the measurement. Mystery shoppers are well trained and know the processes and are therefore able to measure the critical failure points (Anderson, Grove, Lengfelder, & Timothy, 2001). Mystery shoppers know they are going to evaluate an outlet, whereas customers typically only find out afterwards when presented with a survey request (Finn & Kayande, 1999). The mystery shoppers that gather the data are independent, critical, objective and anonymous (Van der wiele, Hesselink, & van Iwaarden, 2005; Hesselink & Van der wiele, 2003). Mystery shopping enables the evaluation of processes instead of outcomes, and importantly this evaluation occurs at the time of the service delivery, i.e. ‘measuring the service as it unfolds’ (Ford, Latham, & Lennox, 2011; Wilson, 1998). According to Wilson (1998) this avoids ‘misremembering’ by the customer, one of the main pitfalls of post-service delivery survey methods.

Although mystery shopping has several advantages and is used for many purposes, several researchers expressed criticism towards the research method.

Criticism on Mystery Shopping

Researchers have highlighted limitations of the mystery shopping method, such as ethical considerations (Erstad, 1998; Ng Kwet Shing & Spence, 2002; Wilson, 2001), generalizability of findings (Finn & Kayande, 1999) and researcher cognition (Morrison, Colman, & Preston, 1997).

Mystery shopping is a clear example of concealed participant observation in a public setting. In terms of ethics, use of deception and observation of people without their knowledge may violate their rights to privacy and freedom from exploitation (Wilson, 2001) According to Wilson (2001) it is critical that service providers are made aware that their employer will be observing them in a concealed manner. Ng Kwet Shing and Spence (2002) took it one step further and stated that mystery shoppers purposefully invite, or provoke, seeking information from not properly authorized personnel.

Besides ethical considerations, another limitation concerns the generalizability of findings in mystery shopping reports. Although Finn and Kayande (1999) found that individual mystery shoppers provided higher quality data than customers, their study demonstrated that assessing outlet performance using mystery shopper reports required more than just two or three mystery shopper visits. They suggested that assessing a single outlet for their interior requires data from about a dozen visits per store, and assessing of service quality requires data from at least 40 to 60 visits per store to create an accurate report.

Finally, Morrison et al. (1997) stated that an important obstacle impeding the reliability of mystery shopping reports is researcher cognition, as the mystery shopper is the research instrument. Since the research instrument is a human being, human error must be considered (Allison, 2009). Human error can occur from memory overload (Boon & Davies, 1993; Morrison et al., 1997). Human error might occur while training or instructing the mystery shopper, during observing (mystery shop visit), or during the reporting of the data. Mystery shopper rely on memory for accuracy of reporting (Morrison et al., 1997), as study materials (e.g. checklists or questionnaires) have to be learned by heart before the mystery shop visit and observations during the visit have to be remembered. However, many researchers in the field of psychology and criminology have showed that human memory is ´bias-sensitive’ (Pezdek, Sperry, & Owens, 2007).

Researcher Cognition

Morrison et al. (1997) introduced challenges mystery shoppers may face when relying on memory for accuracy. Errors can occur at three stages of memory: encoding, storage, or retrieval.

Encoding relies on perceptions of the mystery shopper (Boon & Davies, 1993), and is adjusted by previous service encounters, attitudes, and social pressures (Morrison et al., 1997).

Storage of memory occurs between encoding and retrieval of an event. During this phase the memory is the most sensitive to be influenced by outside influences, affecting the accuracy of the retrieved information (Morrison et al., 1997). Anything that interferes with the storage of memory is liable to lead to incorrect reports affected by the observers' prior expectations rather than objective facts (Baker, 1961). Morrison et al. (1997) stated that observers do not only remember actually perceived events but also observer's expectations. Moreover, these expectancy biases or alterations in memory occur most often in situations in which a large amount of information has to be remembered, for example mystery shopping visit (Macrae, Hewstone, & Griffiths, 1993).

Retrieval is the final action of the memory process, in which memories are retrieved. During the retrieval stage, observers are impressible to alterations in memory, to the point that true memories are replaced by false memories, yet the observer believes the false memory to now be the truth (Leippe, Eisenstadt, Rauch, & Seib. 2004; Leippe, Eisenstadt, Rauch, & Stambush, 2006; Morrison et al., 1997; Pezdek et al., 2007). This implicates that research materials used during mystery shopping visits can affect the accuracy of the reporting, as, for example, type of questioning (open or closed questions) could steer the direction from which memories are retrieved.

This study will focus on effects of forms interference in researcher cognition on accuracy of reporting originating in *storage* and *retrieval* phases of memory, as those phases is most sensitive to be outside influenced (Morrison et. al, 1997).

The effect of forms of interference on researcher cognition could cause difficulties for mystery shoppers to report correctly. They can storage expectations rather than observations (Baker, 1961), or they can retrieve false memories (Leippe et. al, 2004). The extent to which mystery shoppers can report correct should differ based on the number of interferences in research cognition mystery shoppers are confronted with. If mystery shoppers are confronted with multiple forms of researcher cognition bias they are more likely to perform less correct reports than mystery shoppers who are confronted with less forms of researcher cognition bias, or than mystery shopper who are not confronted with any form of researcher cognition bias. Thus, the following hypothesis is given.

> **H1**: Mystery shoppers who are confronted with multiple forms of researcher cognition bias will perform significantly less correct reports during mystery shop visits than mystery shoppers who are confronted with less forms of researcher cognition bias.

The most common forms of interference for mystery shoppers, are the '*expectancy bias*' ("I believe [X], but I observed [Y]") in the storage phase (Hasher & Greenberg, 1977; Holmes, 1972; Taylor, Altman & Sorrentino, 1969) and '*false recollections bias'* ("I remembered [X], but I should have remembered [Y]") (Leippe et al., 2004; Leippe et al., 2006) in the retrieval phase of memory. These forms of interference, or researcher cognition bias, are also known as *disconfirmed expectancies* and *exposure to misinformation*.

Disconfirmed Expectancies

According to the social psychologist Festinger's (1957) theory of cognitive dissonance, cognitive dissonance creates a state of psychological discomfort because the outcome contradicts the expectancy. When an individual receives two ideas which are dissonant, he or she attempts to reduce this mental discomfort by changing of distorting one or both of the ideas to make more consonant. A famous example of cognitive dissonance was a doomsday cult led by Dorothy Martin. Martin claimed to have received messages from aliens forecasting a flood that would end the world. Festinger (1957) found that failure of the prophecy did not break the cult. Instead group members looked for ways to justify their actions and maintain confidence in the cult.

Cognitive dissonance is often paired with disconfirmed expectancy because disconfirmation results in two competing cognitions within individuals. As such, disconfirmed expectancy is often used as a reliable method for inducing cognitive dissonance in experimental designs (Hasher & Greenberg, 1977; Holmes, 1972; Taylor, Altman & Sorrentino, 1969). Generally this is done by introducing an outcome which is dissonant with the participant's established expectations. The expectation is often induced by creating expectancy toward a certain outcome. For example, in the study of Carlsmith and Aronson (1963) participants were asked to taste solutions and rate them on bitterness and sweetness. Following the disconfirmed expectancies participants rated the solutions as less pleasant: sweet solutions were rated less sweet and bitter solutions were rated more bitter.

When recognizing the falsification of an expected event an individual will experience conflicting cognitions, "I believe [X]," but, "I observed [Y]". The individual must either discard the now disconfirmed belief or justify why it has not actually been disconfirmed.

In the case of mystery shopping, mystery shoppers could be confronted with inaccurate study material (e.g. checklist or questionnaire). Mystery shoppers expect to observe certain items, but in practice inaccurate study material could contain items that are not present during the visit which could result in a form of disconfirmed expectancy. For example, a mystery shopper is told to measure service performance of waiters in a restaurant to discover upon arriving it is a ‘self-service restaurant’. The mystery shopper now has to discard the disconfirmed belief that he/she was served by a waiter or justify that he/she was served. Due to psychological discomfort this could cause difficulties for mystery shoppers to provide an accurate report of visits. Therefore the number of correct reports during a mystery shopping visit should differ based on whether mystery shoppers are disconfirmed in their expectancies or not. If mystery shoppers are disconfirmed in their expectancies they are more likely to perform less correct reports than mystery shoppers who are not disconfirmed in their expectancies. Thus, the following is hypothesis is given.

> **H2**: Mystery shoppers who are disconfirmed in their expectancies will perform significantly less correct reports during mystery shop visits than mystery shoppers who are not disconfirmed in their expectancies.

Exposure to Misinformation

Numerous studies have demonstrated that eyewitnesses can be misled by post-event suggestions following an observed event (Loftus, Miller, & Burns, 1978; Pezdek, Finger, & Hodge, 1997; Pezdek, Lam & Sperry, 2009). In most of this research, post-event misinformation has been other-generated, for example, information in an interviewer’s questions about an event that followed viewing an event. Exposure to such a form of misinformation can significantly hamper a person’s ability to provide an accurate report (Lindsay, 1990; Lindsay & Johnson, 1989; Loftus et. al, 1978; Schreiber & Sergent, 1998). Exposure to misinformation can lead people to recall seeing objects that did not appear or occur in the original event (Nourkova, Bernstein, & Loftus, 2004). Researchers have shown that they can also persuade people to recall existence of people or experiences that are completely fictitious (Loftus, 2005). Using various forms of suggestion, researchers have led people to believe they have been hospitalized, attacked by a vicious animal or uncomfortably and repeatedly licked on the ear by a Disney character (Berkowitz, Laney, Morris, Garry, & Loftus, 2008; Heaps & Nash, 2001; Morgan, Southwick, Steffian, Hazlett, & Loftus, 2013).

In case of mystery shopping, mystery shoppers may complete reports and realize not remembering an element and therefore is forced to recreate the event in memory (Morrison et al., 1997). For example, a mystery shopper is told to measure the service performance of waiters in a restaurant to realize after the visit that also aesthetics of the restaurant were to be evaluated. The recreation of the event in memory could cause a form of false recollection, forcing the mystery shopper to create a memory about situations that did not actually occur (Pezdek et al., 2007). This could cause difficulties for the mystery shopper to provide an accurate report. If mystery shoppers are exposed to misinformation, they are more likely to remember items that they never observed and therefore perform less correct reports than mystery shoppers who are not exposed to misinformation. Thus, the following is hypothesized.

> **H3**: Mystery shoppers who are exposed to misinformation will perform significantly less correct reports during mystery shop visits than mystery shoppers who are not exposed to misinformation.

No form of researcher cognition bias

Mystery shoppers who don't experience psychological discomfort of forms of researcher cognition bias should have less difficulty in providing an accurate report of the mystery shop visit, due to the fact that they are confirmed in their expectancy during the visit and are not exposed to a form of misinformation after the visit. If mystery shoppers are not confronted with any form of researcher cognition bias they should perform more correct reports than the mystery shoppers who are confronted with a form of researcher cognition bias. Thus, the following hypothesis is given.

> **H4**: Mystery shoppers who are not confronted with a researcher cognition bias will perform significantly more correct reports during mystery shop visits than mystery shoppers who are confronted with (a) form(s) of researcher cognition bias.

Goal of this study

This study investigates what happens when mystery shoppers are confronted with forms of researcher cognition bias: to what extend influence disconfirmed expectancies the reliability of mystery shopping reports, and/or to what extend influence exposure to misinformation the reliability of mystery shopping report? Are mystery shoppers going to doubt their own observations and produce erroneous reports? Literature suggests they might. This study provides insights into the reliability mystery shopping reports by combining quantitative data of a semi-controlled environment with qualitative data on the motives for behavior.

METHOD

To study effects of disconfirmed expectancies and exposure to misinformation on the reliability of mystery shopper reports this study consisted of two phases. The first phase was a mystery shop experiment. Sixty-three mystery shoppers took part, randomly divided over four conditions in a 2 * 2 field experiment (confirmed expectations versus disconfirmed expectations, and no exposure to misinformation versus exposure to misinformation). With one control group and three experimental groups the goal of this phase was to investigate the effect of the manipulations on the reliability of reporting of mystery shoppers. The second phase was a follow up interview. Nineteen mystery shoppers were randomly selected to participate in the follow up interview, using an evolutionary interview structure. This qualitative method offered explanation and meaning to the findings of the mystery shop visit. Both methods were predetermined, pre-tested (*n* = 15) and rehearsed.

Phase I: Mystery Shop Experiment

The 2 * 2 research design, procedure of the mystery shop experiment, sample of mystery shoppers, checklist used and analyses criteria will be described here.

Design

This study is based on a 2 * 2 design (confirmed expectancies vs. disconfirmed expectancies, and no misinformation vs. misinformation). With one control group and three experimental groups there were four conditions to which mystery shoppers were randomly divided. See Table 1 for an overview of the conditions, their appellation and a brief description of the manipulations. To increase the '*readability*' of this paper, the appellation of the conditions is changed to their practical manipulations name (M) rather than terminology derived from literature (L).

Table 1. ***The 2 * 2 Design with the appellations based on literature (L) and the appellations based on the manipulations (M)***

	L: Disconfirmed Expectancies **M: False Checklist**	L: Confirmed Expectancies **M: True Checklist**
L: Exposure to Misinformation **M: Swap**	Experimental Group (*n* = 16) **M: False Checklist / Swap condition** This group was given the *false checklist* beforehand (which contained four existing and one non existing product item), and was given the *true checklist* upon return (which contained five existing product items). The checklist was '*swapped*' during the visit.	Experimental Group (*n* = 15) **M: True Checklist / Swap condition** This group was given the *true checklist* beforehand (which contained five existing product items), and was given the *false checklist* upon return (which contained four existing and one non existing product item). The checklist was '*swapped*' during the visit.
L: No Exposure to Misinformation **M: No Swap**	Experimental Group (*n* = 15) **M: False Checklist / No Swap condition** This group was given the *false checklist* (which contained four existing and one non existing product item) beforehand and upon return. The checklist was '*not swapped*' during the visit.	Control group (*n* = 17) **M: True Checklist / No Swap condition** This group was given the *true checklist* (which contained five existing product items) beforehand and upon return. The checklist was '*not swapped*' during the visit.

Procedure

The mystery shoppers were welcomed by the research leader in an office nearby the location of the visit (local supermarket). They were told to perform two mystery shop visits (*trial* and *actual* visit) commissioned by the headquarter of the supermarket.

Trial visit

Before the *actual* visit, the mystery shoppers participated in a *trial* visit. The mystery shoppers were provided with an instruction document containing: an act of confidentiality, a protocol and a checklist (see Appendix III: Instruction document). The mystery shoppers had to sign the act of confidentiality read the protocol and learn the items of the checklist by heart. The main task of the *trail* visit was to evaluate service performance of the butchery within the supermarket. After the instructions the mystery shopper had to go undercover to evaluate the items of the checklist. By doing so they gained experience with the mystery shop method, the supermarket, and the materials used. No manipulations took place during the *trial* visit. This was essential for the effect of the manipulations in the *actual visit* because the mystery shopper were now primed on the trustworthiness of the research leader and the materials used. To increase the realism of the tasks, all the materials were printed in the design of the supermarket.

After the *trial* visit, the mystery shoppers returned to the office nearby where the filled out the checklist. They had the possibility to ask questions about the method and received instructions for the *actual* visit.

Actual visit

For the *actual* visit the mystery shoppers were instructed, in the office nearby, with the fake research goal to perform a quality assessment of the interior of the supermarket in which they were not allowed to engage in interactions with the supermarket staff. The mystery shoppers were again provided with an instruction document containing an act of confidentiality, a protocol and a checklist. The mystery shopper had to sign the act of confidentiality, read the protocol and learn the items of the checklist by heart. The main task of the *actual* visit was to note the exact price of five product items. After the instructions the mystery shoppers had to go undercover to evaluate the items of the checklist. During the *actual* visit manipulations occurred for the experimental groups (see Checklist – Manipulations).

After the *actual* visit, the mystery shoppers returned to the office nearby to fill out the checklist. They had the possibility to speak up to the research leader, or ask questions about experienced difficulties. Furthermore a Q & A list was made in order to answer any questions of mystery shoppers. The experiment ranged in length from 35 min to 1 h 20 min.

Mystery Shoppers

In total, 64 mystery shoppers participated in this phase. All mystery shoppers were Dutch speaking students, and were recruited on a voluntary basis, with the incentive of gaining two '*study-credits*'. One mystery shopper was excluded from the sample due to the fact that he could not properly understand the Dutch language, which was required for this study. The gross sample was *N* = 63.

In total, 22 men and 41 women participated (average age = 21.24). Most of the mystery shoppers (*n* = 47) had no mystery shopping experience, *n* = 8 had very little experience, *n* = 2 had somewhat experience, *n* = 2 had much experience and *n* = 4 had a great deal of experience with mystery shopping.

Gender (χ^2(3, N = 63) = 4.072, *p* = .254), age (χ^2(33, N = 63) = 36.161, *p* = .323) and experience (χ^2(12, N = 63) = 9.529, *p* = .657) were randomly divided among the four conditions. The conditions were randomly divided over the days of the week (χ^2(6, N = 63) = 11,588, *p* = .710), the time of the day (χ^2(6, N = 63) = 11,371, *p* = .726), the level of activity in the supermarket (χ^2(3, N = 63) = 3,582, *p* = .733) and the evaluation grade of the supermarket (F(3, 59) = .7, *p* = .556). See Table 2 for an overview of the mystery shoppers that participated in this phase.

Table 2: ***Overview of Mystery Shoppers and Visits***

	True Checklist No Swap	**False Checklist No Swap**	**True Checklist Swap**	**False Checklist Swap**	**Total**
Gender					
Man	4 (23.5%)	8 (53.3%)	6 (40%)	4 (25%)	**22 (34.9%)**
Women	13 (76.5%)	7 (46.7%)	9 (60%)	12 (75%)	**41 (65.1%)**
Average Age	20.94	22.6	21.13	20.38	**21.24**
Level of experience with mystery shopping					
None	14 (29.8%)	9 (19.1%)	10 (21.3%)	14 (29.8%)	**47 (74.6%)**
Very Little	1 (12.5%)	3 (37.5%)	3 (37.5%)	1 (12.5%)	**8 (12.7%)**
Somewhat	0 (0%)	1 (50%)	1 (50%)	0 (0%)	**2 (3.1%)**
Much	0 (0%)	1 (50%)	0 (0%)	1 (50%)	**2 (3.1%)**
Great Deal	2 (50%)	1 (25%)	1 (25%)	0 (0%)	**4 (6.5%)**
Day of week					
Monday	5 (50%)	2 (20%)	2 (20%)	1 (10%)	**10 (15.9%)**
Tuesday	0 (0%)	1 (50%)	0 (0%)	1 (50%)	**2 (3.1%)**
Wednesday	4 (22.2%)	4 (22.2%)	4 (22.2%)	6 (33.3%)	**18 (28.6%)**
Thursday	3 (23.1%)	2 (15.4%)	5 (38.5%)	3 (23.1%)	**13 (20.6%)**
Friday	5 (26.3%)	6 (31.6%)	3 (15.8%)	5 (26.3%)	**19 (30.1%)**
Sunday	0 (0%)	0 (0%)	1 (100%)	0 (0%)	**1 (1.5%)**
Time of day					
8-10am	1 (16.7%)	2 (33.3%)	1 (16.7%)	2 (33.3%)	**6 (9.5%)**
10-12am	8 (44.4%)	3 (16.7%)	4 (22.2%)	3 (16.7%)	**18 (28.6%)**
12am-14pm	1 (10%)	2 (20%)	3 (30%)	4 (40%)	**10 (15.9%)**
14-16pm	3 (37.5%)	2 (25%)	1 (12.5%)	2 (25%)	**8 (12.7%)**
16-18pm	4 (25%)	5 (31.2%)	3 (18.8%)	4 (25%)	**16 (25.4%)**
18-20pm	0 (0%)	1 (20%)	3 (60%)	1 (20%)	**5 (7.9%)**
Level of activity in supermarket I					
Calm	5 (20.8%)	7 (29.2%)	6 (25%)	6 (25%)	**24 (38.1%)**
Normal	9 (29%)	5 (16.1%)	8 (25.8%)	9 (29%)	**31 (49.2%)**
Busy	3 (37.5%)	3 (37.5%)	1 (12.5%)	1 (12.5%)	**8 (12.7%)**
Total	17 (27%)	15 (23.8%)	15 (23.8%)	16 (25.4%)	**63 (100%)**
Level of activity in supermarket II [open]					
N cash desks total	4.59	4.60	4.80	4.44	**4.6**
N cash desks open	2.12^{a}	1.93^{a}	1.27	1.88^{a}	**1.81**
N customers in front	1.82^{ab}	2.20^{a}	1.07^{b}	1.94^{ab}	**1.76**
N customers behind	1.24	1.13	1.80	1.50	**1.41**
Familiarity with products [1-5]					
Filler product A	1.76	2.07	2.60	1.81	**2.05**
Filler product B	1.65	2.13	1.80	1.63	**1.79**
Target product	1.94^{a}	3.27^{b}	2.40^{ab}	2.19^{a}	**2.43**
Filler product C	2.41	2.33	2.53	2.06	**2.33**
Filler product D	1.94	2.00	2.07	1.88	**1.97**
Evaluation grade of supermarket [1-10]					
Evaluation grade	7.41	7.4	7.33	7.06	**7.3**

* $p < .05$ ** $p < .01$ (Chi Square)

[ab] Means followed by the same letter within columns were not significantly different ($p < 0.05$) (One Way ANOVA Post Hoc LSD)

Mystery shoppers speaking up to research leader

In total, sixteen mystery shoppers took initiative and spoke up to the research leader about perceived difficulties, biases, and/or manipulations concerning the study, material and/or method. They were equally divided among gender (eight men and eight women) and randomly divided over age ((F(3, 12) = .504, *p* = .687) with an average age of 21.44. They were self-selected from the total of N = 63, with a random level of experience with the mystery shopping method among them ((χ^2(9, N = 63) = 11.905, *p* = .219). See Table 3 for an overview of mystery shoppers who spoke up to the research leader.

Table 3: ***Overview of Mystery Shopper who spoke up to the research leader***

	True Checklist No Swap	**False Checklist No Swap**	**True Checklist Swap**	**False Checklist Swap**	**Total**
Gender					
Man	0 (0%)	5 (62.5%)	2 (25%)	1 (12.5%)	**8 (50%)**
Women	2 (25%)	2 (25%)	0 (0%)	4 (50%)	**8 (50%)**
Average Age	20	22.3	21	21	**21.44**
Level of Experience					
None	2 (16.7%)	4 (33.3%)	1 (8.3%)	5 (41.7%)	**12 (75%)**
Very Little	0 (0%)	2 (100%)	0 (0%)	0 (0%)	**2 (12.5%)**
Somewhat	0 (0%)	0 (0%)	1 (100%)	0 (0%)	**1 (6.25%)**
Great Deal	0 (0%)	1 (100%)	0 (0%)	0 (0%)	**1 (6.25%)**
Total	2 (12.5%)	7 (43.8%)	2 (12.5%)	5 (31.2%)	**16 (100%)**

Chi-square analyses with * p < .05,** p < 0.01

[ab] Means followed by the same letter within columns were not significantly different (*p* < 0.05) (One Way ANOVA Post Hoc LSD)

Dependent variable – Checklist

Based on items from published scales (Dawes & Sharp, 2000; Finn, 2001; Finn & Kayandé, 1999; Gosselt, van Hoof, de Jong & Prinsen, 2007; Hesselink & van der Wiele, 2003; Kocevar-Weidinger, Benjes-Small, Ackermann & Kinman, 2009; Struyk, 2012; Tarantola, Vicard & Ntzoufras, 2012; Van der Wiele, Hesselink & van Iwaarden, 2005;) two *genuine looking* checklists consisting of thirty-two items divided over five categories were developed to measure the effects of disconfirmed expectancies and exposure to misinformation on mystery shopping reports (see Appendix II: Checklist). A *true version* and a *false version*:

The items of the checklists were identical in both versions; except the target product item in '*Interior*' category, that is where manipulations occurred. The '*Interior'* category consisted in both versions of four filler products and one target product. All filler products were existing product items and did not differ per version. In the *true* checklist the target product was an existing product item but in the *false* checklist the target product was a non-existing product item, which meant that the product literally did not exist. Table 4 shows product items on the checklist before and after the visit for the *true* and *false* versions.

Table 4: ***Product Items in Interior Category on checklist before and after visit***

	True Checklist No Swap	**True Checklist Swap**	**False Checklist No Swap**	**False Checklist Swap**
	Filler Product A	Filler Product A	Filler Product A	Filler Product A
	Filler Product B	Filler Product B	Filler Product B	Filler Product B
Before Visit	***Existing Target Product***	***Existing Target Product***	***Non-Existing Target Product***	***Non-Existing Target Product***
	Filler Product C	Filler Product C	Filler Product C	Filler Product C
	Filler Product D	Filler Product D	Filler Product D	Filler Product D
	Filler Product A	Filler Product A	Filler Product A	Filler Product A
	Filler Product B	Filler Product B	Filler Product B	Filler Product B
After Visit	***Existing Target Product***	***Non- Existing Target Product***	***Non-Existing Target Product***	***Existing Target Product***
	Filler Product C	Filler Product C	Filler Product C	Filler Product C
	Filler Product D	Filler Product D	Filler Product D	Filler Product D

Manipulation: False checklist

In order to simulate disconfirmed expectancies the mystery shoppers were provided with the *false version* of the checklist containing a non-existing product item (ratio 4:1 to existing product items). Thus, a mystery shopper was provided with the *false* checklist which contained a product item that did not exist (non-existing product item) and thus was not available at the supermarket, but was told to be available at store.

Manipulation: Swap

In order to simulate the exposure to misinformation the provided checklist beforehand was 'swapped' (changed/switched) to the different version of the checklist during the visit of the mystery shoppers. Thus, a mystery shopper was provided with the *true* version of the checklist before the visit and is provided with the *false* version of the checklist upon return (or vice versa). The mystery shopper had to remember the items of the checklist by heart and leave the checklist at the office during the visit. To increase realism, the act of confidentiality that was signed by the mystery shopper before the visit was removed from the checklist and 're-stapled' to the different version of the checklist.

Analyses

The results of the mystery shopping experiment were analyzed with IBM SPSS Statistics 20.

Coding (in-)correct reports in the '*Interior*' category: The main task of the mystery shoppers was to note the exact correct prices of five products. See Table 5 for an overview of the coding of (in-)correct report per product type.

1) *Filler product*: If a mystery shopper noted the exact correct price of a filler product that was coded as a correct report. If they noted an incorrect price of a filler product that was coded as an incorrect report. If a mystery shopper spoke up to the research leader about experienced difficulties that was also coded as a correct report.
2) *Target product*: In the control group it was possible to find the target product. Therefore a correct price was coded as a correct report and an incorrect price was coded as incorrect report. If a mystery shopper spoke up to the research leader about experienced difficulties that was also coded as a correct report.

In the experimental conditions it was not possible to note the correct price of the target product. The target product on the checklist either did not exist and thus was not available in store (*false checklist*) and/or the checklist was '*swapped*' during the visit confronting the mystery shopper with a different target product on the checklist upon return from their visit and therefore the mystery shopper could never have seen that product item. If the mystery shopper spoke up to the research leader about experienced difficulties that was coded as a correct report. All other reports were noted as incorrect.

Table 5: ***Overview of Coding (In-)Correct Reports per Product Type***

	Correct Reports	**Incorrect Reports**
Filler Product	▪ Correct price, in all conditions ▪ Spoke up about experienced difficulties	▪ Incorrect price, in all conditions
Target Product	▪ Control condition: exact correct price or spoke up to the research leader about experienced difficulties. ▪ Disconfirmed expectancies: Spoke up to the research leader about experienced difficulties (non-existing target product). Either face-to-face or on the checklist. ▪ Exposure to misinformation: Spoke up to the research leader about experienced difficulties ('swap'). Either face-to-face or on the checklist.	▪ Incorrect price, in all conditions ▪ Disconfirmed expectancies: Filling out a price for the target product. Not speaking up to the research leader about experienced difficulties. ▪ Exposure to misinformation: Filling out a price for the target product. Not speaking up to the research leader about experienced difficulties.

Phase II: Follow up Interview

The procedure, sample, coding scheme and analyses used will be described here.

Procedure

After '*Phase I: Mystery Shop Experiment'* mystery shoppers could be invited to participate in the follow up interview. The goal of the follow up interview was to gain explanation and meaning to findings of mystery shop visits and the effect of researcher cognition bias. The follow up interview had an evolutionary structure, containing two standardized questions. Mystery shoppers who finished the mystery shop experiment (*Phase I*) did not know the true purpose of the experiment. Fearing that future mystery shoppers would know on forehand that they could be manipulated the true purpose of the experiment was kept secret. Based on findings from the pretest the research leader wrote along during the interview instead of recording it to diminish suspicion by the mystery shoppers. All mystery shoppers were debriefed by telling the true purpose of the experiment.

Sample

The sample of the follow up interviews consisted of nineteen mystery shoppers, of which 42.1% man (n = 8) en 57.9% women (n = 11) with an average age of 22 (F(3, 15) = .731, p = .507). All were Dutch speaking students who had participated in '*Phase I: Mystery Shop Experiment*'. They were randomly selected from a total of N = 63 between the conditions ((χ^2(3, N = 63) = .403, p = .940), with a random level of experience with the mystery shopping method among them ((χ^2(9, N = 63) = 9.014, p = .436). The interviews ranged in length from approximately 3 and half minute to 6 minutes.

Due to self-selecting of mystery shoppers who spoke up to the research leader after the *actual visit* five mystery shoppers who spoke up where also interviewed in the follow up interview. See Table 6 for an overview of the follow up interview participants.

Table 6: ***Overview of Follow up Interview Participants***

	True Checklist No Swap	**False Checklist No Swap**	**True Checklist Swap**	**False Checklist Swap**	**Total**
Gender					
Man	2 (25%)	2 (25%)	2 (25%)	2 (25%)	**8 (42.1%)**
Women	2 (18.2%)	2 (18.2%)	4 (36.4%)	3 (27.3%)	**11 (57.9%)**
Average Age	23.5	22.5	20.7	22	**22**
Level of Experience					
None	3 (20%)	3 (20%)	5 (33.3%)	4 (26.7%)	**15 (78.9%)**
Very Little	0 (0%)	0 (0%)	1 (50%)	1 (50%)	**2 (10.5%)**
Somewhat	0 (0%)	1 (100%)	0 (0%)	0 (0%)	**1 (5.3%)**
Great Deal	1 (100%)	0 (0%)	0 (0%)	0 (0%)	**1 (5.3%)**
Total	**4 (21.1%)**	**4 (21.1%)**	**6 (31.6%)**	**5 (26.3%)**	**19 (100%)**

Chi-square analyses with * p < .05,** p < 0.01

^{ab} Means followed by the same letter within columns were not significantly different (p < 0.05) (One Way ANOVA Post Hoc LSD)

Coding Scheme Interviews and Analyses

Answers of mystery shoppers in the follow up interviews were sentence-based coded into six categories, which were divided in twenty codes. Answers given could have multiple codes, which resulted in a total of 165 remarks (see Table 7). The findings of the interviews were analyzed in the analyses program ‘Atlas.ti’.

Table 7: ***Categories and Codes used in Follow up Interview***

Categories	**Codes**	***n* of remarks**
‘Interior’ Products (*n* = 10 remarks)	Target product	**3**
	Filler product A	**3**
	Filler product B	**2**
	Filler product C	**2**
	Filler product D	**0**
Visit Experiences (*n* = 44 remarks)	‘Findability’ Products	**13**
	Remembering Products	**12**
	Swap Checklist	**9**
	Remembering Prices	**6**
	Wrong. Remembered	**4**
Reactions (*n* = 18 remarks)	Excuses	**11**
	Motivation	**5**
	Mental Support Tricks	**2**
Length (*n* = 6 remarks)	Time Visit	**4**
	Time Checklist	**2**
Attitudes (*n* = 39 remarks)	Positive	**22**
	Negative	**17**
Study (*n* = 48 remarks)	Study Difficulty	**26**
	Study General	**16**
	Study Material	**6**
Total		**165**

RESULTS

First results from quantitative data (phase I: *mystery shopping experiment*) and finally results from qualitative data (phase II: *follow up interviews*) will be presented.

Quantitative results - Mystery Shopping Experiment

The statistical analysis of all filled out checklists shows an average of 3.33 out of five correct reports per mystery shopper (see also Table 8), which represents an average of 1.67 incorrect reports per mystery shopper. Out of the 315 product items observed, 217 times the mystery shoppers reported correct. This involves noting the exact correct price for filler products, or admitting an error, noting the correct price or noting the manipulation for the target product. This leaves 98 incorrect reports by mystery shoppers. This involves noting the incorrect price for filler products and target products, but also filling out the target product when mystery shopper could not find the target product in the *false checklist* conditions or when the mystery shoppers could not have known that the target product was going to appear, due to the swap, on the checklist in the *swapped checklist* condition.

Table 8: ***Overview of correct reports*** **(Y)*****, incorrect reports (N) and average number of correct reports (0-5) per condition and product type (target / filler)***

			N of correct reports										**Average** **(out of five)**
			Target product**		Filler products [A-B-C-D]								
					A		B		C		D		
			Y	N	Y	N	Y	N	Y	N	Y	N	
Condition** **N = 63**	False Checklist	Swap (*n* = 16)	3	13	13	3	10	6	11	5	9	7	**2,88**^{a}
		No Swap (*n* = 15)	5	10	14	1	5	10	8	7	12	3	**3,20**^{a}
	True Checklist	Swap (*n* = 15)	1	14	13	2	10	5	12	3	12	3	**2,93**^{a}
		No Swap (*n* = 17)	15	2	15	2	12	5	15	2	15	2	**4,24**
Total			24	39	55	8	38	25	53	10	46	17	**315/3,33**

Chi-square analyses with * p < .05,** p < 0.01

^{ab} Means followed by the same letter within columns were not significantly different (*p* < 0.05) (One Way ANOVA Post Hoc LSD)

First the results *between* the conditions and finally the results *within* the conditions will be presented.

Overview of (in)correct reports between the four conditions

Table 8 shows an overview of the number of (in)correct reports per condition and per product type. A Chi-Square test of *Condition * N of Correct Reports* (χ^2(3, N = 63) = 30,851, *p* = .009) showed that there was a significant difference for the number of correct reports between the conditions. Experimental groups showed no significant differences between each other for the number of correct reports *False Checklist / No Swap * True Checklist / Swap* (χ^2(4, N = 63) = 2.5, *p* = .645), *False Checklist / Swap * True Checklist* / *Swap* (χ^2(5, N = 63) = 6.2, *p* = .282) and *False Checklist* / *No Swap * False Checklist* / *Swap* (χ^2(5, N = 63) = 2.37, *p* = .796). These results rejected hypothesis 1 that mystery shoppers who are confronted with multiple forms of researcher cognition bias will perform significantly less correct reports during mystery shop visits than mystery shoppers who are confronted with less forms of researcher cognition bias (one or none). However, when comparing the *false checklist* / *swap* condition with the control group (*true checklist* / *no swap*) a Chi-square test (χ^2(5, N = 63) = 12.945, *p* = .024) showed that there was a significant difference in the number of correct reports between the two conditions. Even though these results support a part of hypothesis 1 the hypothesis remains rejected due to the fact that the *false checklist* / *swap* condition showed no significant difference in the number of correct reports in comparison to other experimental groups, as mentioned above. Thus, hypothesis 1 remains rejected.

A Chi-square test of *False Checklist* / *No Swap* True Checklist / No Swap* (χ^2(4, N = 63) = 10.647, *p* = .031) showed a significant difference in the number of correct reports between mystery shoppers who were disconfirmed in their expectancies and mystery shopper who were not disconfirmed in their expectancies. These results support hypothesis 2.

A Chi-square test of *True Checklist / Swap * True Checklist / No Swap* (χ^2(3, N = 63)

= 9.73, *p* = .021) showed that there was a significant difference in the number of correct reports between mystery shoppers who exposed to misinformation and mystery shoppers who were not exposed to misinformation. These results support hypothesis 3.

Results of Chi-square tests above showed that the *True Checklist / No Swap* condition (control group) scored significantly different from all other conditions. These findings were supported by an *One Way ANOVA Test* with a value of F(3, 59) = 4,534, *p* = .006 which showed that there was a significant difference in the average number of correct reports between conditions. Post hoc comparisons using the LSD test indicated that mean scores for the *true checklist* / *no swap* condition (M = 4.24, SD = 1.09) was significantly different from *experimental* conditions (M = 3.20, SD = .941 / M = 2.93, SD = 1.22 / M = 2.88, SD = 1.50). With an average of 4.24 correct reports out of 5, this condition showed an 85% correct report rate. These results support hypothesis 4 that mystery shoppers who are not confronted with researcher cognition bias will perform significantly more correct reports during mystery shop visits than mystery shoppers who are confronted with (a) form(s) of researcher cognition bias.

The next step was to determine where the significant difference between conditions originated from. Four Chi-square tests of *Condition * N of Correct Reports Filler Products [A-B-C-D]* showed no significant differences in the number of correct reports between the conditions for filler products. This means that the significant difference for the number of correct reports between conditions did not originated from *filler products*.

A Chi-square test of *Condition * N of Correct Reports Target Product* (χ^2(3, N = 63) = 27,089, *p* = .001) showed that there was a significant difference in the number of correct reports between conditions for the target product. This means that the significant difference for the number of correct report between conditions could originate from the *target product*.

Four more Chi-square tests were conducted to investigate where the significant difference of the *target product* precisely occurred. *No Swap * N of Correct Reports Target*

Product showed a χ^2(1, N = 63) = 10,2418, *p* = .002, *Swap * N of Correct Reports Target Product* showed a χ^2(1, N = 63) = 1,006, *p* = .325, *True Checklist * N of Correct Reports Target Product* showed a χ^2(1, N = 63) = 21,208, *p* = .001, and *False Checklist * N of Correct Reports Target Product* showed a χ^2(1, N = 63) = .860, *p* = .303. These analyses showed that the significant difference for the number of correct reports for the target product between conditions originated from the *True Checklist / No Swap* condition. These results support hypothesis 4.

In the following paragraphs the results will described per and within conditions, starting with the *false checklist / swapped* condition.

False Checklist / Swapped

Sixteen mystery shoppers were confronted with a false checklist that was swapped to the different version during their visit. Out of the 80 reports, they reported 47 times (58.7%) correct which leaves 33 times (41.3%) in which they reported incorrect. On average, mystery shoppers reported 2.88 (57.6%) out of five items correct which leaves 2.12 (42.4%) in which they reported incorrect. A Cochran's Q Test (*Q* (4, N = 16) = 4, *p* = .001) showed a significant difference between the products for the number of correct reports. Post hoc comparison using the McNemar test revealed that the number of correct reports between the products significantly differed (see Table 9). The target product showed significant less correct reports in comparison to filler product [A] and [C]. However, filler product [B] and [D] showed no significant difference in comparison to the target product or filler products [A] and [C].

Table 9: ***Overview of (In-)Correct reports per product for the False Checklist / Swapped condition(n = 16)***

Products	**Correct Report**	**Incorrect Report**	**Total**
Filler Product A[b]	13 (81.2%)	3 (19.8%)	16 (20%)
Filler Product B[ab]	10 (62.2%)	6 (37.3%)	16 (20%)
Target Product[a]	3 (19.8%)	13 (81.2%)	16 (20%)
Filler Product C[b]	11 (62.2%)	5 (37.3%)	16 (20%)
Filler Product D[ab]	9 (68.7%)	7 (31.3%)	16 (20%)
Total	**47 (58.7%)**	**33 (41.3%)**	**80 (100%)**
Average	2.88 (57.6%)	2.12 (42.4%)	5 (100%)

[ab] Products followed by the same letter within columns were not significantly different (*p* < 0.05) (Cochran's Q Post hoc McNemar test)

False Checklist / No Swap (Disconfirmed Expectancies)

Fifteen mystery shoppers were confronted with a false checklist that was not swapped during their visit. Out of the 75 reports, they reported 44 times (58.7%) correct which leaves 31 times (41.3%) in which they reported incorrect. On average, mystery shoppers reported 2.88 (57.6%) out of five items correct which leaves 2.12 (42.4%) in which they reported incorrect. A Cochran's Q Test (*Q* (4, N = 15) = 19.09, *p* = .001) showed a significant difference between the products for the number of correct reports. Post hoc comparison using the McNemar test revealed that the number of correct report between the products significantly differed (see Table 10). Filler product [A] and [D] were significantly more correctly reported in comparison to filler product [B] and the target product. However, filler product [C] showed no significant difference with either of the "two groups".

Table 10: ***Overview of (In-)Correct reports per product for the False Checklist / No Swap condition(n = 15)***

Products	Correct Report	Incorrect Report	Total
Filler Product A^{b}	14 (93.3%)	1 (6.7%)	15 (20%)
Filler Product B^{a}	5 (33.3%)	10 (66.7%)	15 (20%)
Target Product^{a}	5 (33.3%)	10 (66.7%)	15 (20%)
Filler Product C^{ab}	8 (53.3%)	7 (46.7%)	15 (20%)
Filler Product D^{b}	12 (80%)	3 (20%)	15 (20%)
Total	**44 (58.7%)**	**31 (41.3%)**	**75 (100%)**
Average	3.2 (64%)	1.8 (36%)	5 (100%)

^{ab} Products followed by the same letter within columns were not significantly different (*p* < 0.05) (Cochran's Q Post hoc McNemar test)

True Checklist / Swap (Exposure to Misinformation)

Fifteen mystery shoppers were confronted with a true checklist that was not swapped during their visit. Out of the 75 reports, they reported 48 times (64%) correct which leaves 27 times (36%) in which they reported incorrect. On average, mystery shoppers reported 2.93 (58.6%) out of five items correct which leaves 2.07 (41.4%) in which they reported incorrect. A Cochran's Q Test (*Q* (4, N = 15) = 26.27, *p* = .001) showed a significant difference between the products for the number of correct reports. Post hoc comparison using the McNemar test revealed that the number of correct reports for the target product significantly differed from all the filler products. The filler products showed no significant difference among each other for the number of correct reports (see Table 11).

Table 11: ***Overview of (In-)Correct reports per product for the True Checklist / Swap condition(n = 15)***

Products	**Correct Report**	**Incorrect Report**	**Total**
Filler Product A[a]	13 (86.7%)	2 (13.3%)	15 (20%)
Filler Product B[a]	10 (66.7%)	5 (33.3%)	15 (20%)
Target Product	1 (6.7%)	14 (93.3%)	15 (20%)
Filler Product C[a]	12 (80%)	3 (20%)	15 (20%)
Filler Product D[a]	12 (80%)	3 (20%)	15 (20%)
Total	**48 (64%)**	**27 (36%)**	**75 (100%)**
Average	2.93 (58.6%)	2.07 (41.4%)	5 (100%)

[ab] Products followed by the same letter within columns were not significantly different ($p < 0.05$) (Cochran's Q Post hoc McNemar test)

True Checklist / No Swap (control group)

Seventeen mystery shoppers were confronted with a true checklist that was not swapped during their visit, and thus were not confronted a form of research cognition bias. Out of the 85 reports, they reported 72 times (84.7%) correct which leaves 13 times (13.3%) in which they reported incorrect. On average, mystery shoppers reported 4.24 (84.7%) out of five items correct which leaves .76 (13.3%) in which they reported incorrect. A Cochran's Q Test (Q (4, N = 17) = 4, p = .406) showed no significant difference between the products for the number of correct reports (see Table 12).

Table 12: ***Overview of (In-)Correct reports per product for the True Checklist / No Swap condition (n = 17)***

Products	**Correct Report**	**Incorrect Report**	**Total**
Filler Product A	15 (88.2%)	2 (11.8%)	17 (20%)
Filler Product B	12 (70.6%)	5 (29.4%)	17 (20%)
Target Product	15 (88.2%)	2 (11.8%)	17 (20%)
Filler Product C	15 (88.2%)	2 (11.8%)	17 (20%)
Filler Product D	15 (88.2%)	2 (11.8%)	17 (20%)
Total	**72 (84.7%)**	**13 (13.3%)**	**85 (100%)**
Average	4.24 (84.8%)	0.76 (15.2%)	5 (100%)

[ab] Products followed by the same letter within columns were not significantly different ($p < 0.05$) (Cochran's Q Post hoc McNemar test)

Other results

An *One Way ANOVA Test* with a value of F(3, 41) = 3.26, *p* = .031 showed a significant difference in the average time it took mystery shoppers to complete the visit between the conditions. Post hoc comparisons using the LSD test indicated that the mean score for the *true checklist conditions* (M = 10.83, SD = 2.72 & M = 9.31, SD = 1.87) was significantly different from the *false checklist conditions* (M = 12.83, SD = 4.98 & M = 14.08, SD = 3.84).

An *One Way ANOVA Test* with a value of F(3, 43) = 5.50, *p* = .003 showed a significant difference in the average time it took mystery shoppers to fill out the checklist between the conditions. Post hoc comparisons using the LSD test indicated that the mean score for the *no swap conditions* (M = 5.15, SD = .98 & M = 6.17, SD = 1.69) was significantly different from the *swap conditions* (M = 7.80, SD = 1.54 & M = 7.75, SD = 2.86). See Table 13 for an overview of the average times it took mystery shoppers to complete the visit and the checklist per condition.

Table 13: ***Overview of the time results per condition***

	True Checklist No Swap	**True Checklist Swap**	**False Checklist No Swap**	**False Checklist Swap**	**Average**
Visit	10:50[a]	12:50[b]	9:31[a]	14:05[b]	**11:59**
Fill out checklist	05:09[a]	06:10[a]	07:48[b]	07:45[b]	**06:38**

[ab] Means followed by the same letter within columns were not significantly different (p < 0.05) (One Way ANOVA Post Hoc LSD)

Summary of Hypotheses

Table 14: ***Summary of Hypotheses, Tests, and Outcomes***

Hypothesis	Tests	Outcome
H1: Mystery shoppers who are confronted with multiple forms of researcher cognition bias will perform significantly less correct reports during mystery shop visits than mystery shoppers who are confronted with less forms of researcher cognition bias (one or none).	Chi-Square Analysis, One Way ANOVA Post Hoc LSD	**Rejected**
H2: Mystery shoppers who are disconfirmed in their expectancies will perform significantly less correct reports during mystery shop visits than mystery shoppers who are not disconfirmed in their expectancies.	Chi-Square Analysis	**Accepted**
H3: Mystery shoppers who are exposed to misinformation will perform significantly less correct reports during mystery shop visits than mystery shoppers who are not exposed to misinformation.	Chi-Square Analysis	**Accepted**
H4: Mystery shoppers who are not confronted with researcher cognition bias will perform significantly more correct reports during mystery shop visits than mystery shoppers who are confronted with (a) form(s) of researcher cognition bias.	Chi-Square Analysis, One Way ANOVA Post Hoc LSD	**Accepted**

Qualitative results – Follow up Interview

The results of the follow up interviews will be presented per condition.

False Checklist / Swap

In total, 16 mystery shoppers were provided with a false checklist which contained non-existing target products (e.g. disconfirmed in their expectancies) and were confronted with a different version of the checklist after their visit (e.g. exposure to misinformation). Of the sixteen, 18.7% (*n* = 3) was invited to participate in the follow up interview. In total, 31.3% (*n* = 5) took initiative and spoke up to the research leader about perceived difficulties. This leaves 68.7% (*n* = 11) of the mystery shoppers who didn't initiate any form of interaction with the research leader even though they were provided with a checklist which contained non-existing target products and were confronted with a different version of the checklist after their visit.

All three mystery shoppers who were invited for the follow up interview stated that they experienced difficulty during or after the visit:

- "When I returned I noticed that I remembered the wrong products. I was really starting to doubt myself" (resp. #4, male)
- "Upon return I noticed I remembered the wrong Taksi [target product]" (resp. #8, female)
- "I don't know how you did it but the checklist was different" (resp. #18, female)
- "I noted the prices of different (the wrong) products" (resp. #18, female)

Two mystery shoppers who were invited for the follow up interview after they spoke up to the research leader stated that they remembered the wrong products.

- "I guess I remembered it wrong, but I thought there were other products on the list" (resp. #12, female)

One of five mystery shoppers who spoke up to the research leader thought she did something wrong and gave an explanation why she provided a possible incorrect report:

- "I just filled something out, otherwise it looked like I couldn't remember anything" (resp. #12, female)

Another mystery shopper who spoke up to the research leader gave a possible reason for incorrect reporting. He stated that the research leader is dependent on the willingness of the mystery shopper to speak up:

- “If I would not have said this, you would never know that I remembered the wrong products” (resp. #15, male)

False Checklist / No Swap (disconfirmed expectancies)

In total, 15 mystery shoppers were provided with a false checklist which contained a non-existing target product (e.g. disconfirmed in their expectancies). Of the fifteen, 20% (*n* = 3) was invited to participate in the follow up interview. In total, 46.7% (*n* = 7) took initiative and spoke up to the researcher about perceived difficulties. This leaves 53.3% (*n* = 8) of the mystery shoppers who didn’t initiate any form of interaction with the research even though their checklist contained non-existing target products.

One of three mystery shoppers who were invited to participate in the follow up interview stated that the non-existing target product truly was non-existing:

- “The Taksi [Target product] was not there, or it was in a completely illogic place” (resp. #11, female)

Two mystery shoppers stated that the non-existing target product was findable, even though that product did not exist:

- “It was quite a search but eventually I found them all” (resp. #3, female)
- “I found the products rather quick” (resp. #7, male)

One of seven mystery shoppers who spoke up to the research leader was invited to participate in the follow up interview. He questioned the ‘*findability*’ and existence of the non-existing target product and gave a remark about the ‘findability’ of filler product [D]:

- “I couldn’t find the Taksi [Target product]” (resp. #16, male)
- “Does that Taksi [Target product] even exist?” (resp. #16, male)
- “Campina [Filler product D] was quite a search!”(resp. #16, male)

True Checklist / Swap (exposure to misinformation)

In total, 15 mystery shoppers were confronted with a different version of the checklist after their visit (e.g. exposure to misinformation). Of the fifteen, 33.3% (*n* = 5) was invited to participate in the follow up interview. In total, 13.3% (*n* = 2) took initiative and spoke up to the research leader about perceived difficulties. This leaves 86.7% (*n* = 13) of the mystery shoppers who didn't initiate any form of interaction with the research leader even though they were confronted with a different version of the checklist after their visit.

Two of five mystery shoppers who were invited to participate in the follow up interview mentioned a change in product items on their checklist after their visit:

- ”I guess I remembered it wrong, but I thought there were other products on the checklist” (resp. #2, female)
- “I don't know why, but I remembered the wrong products” (resp. #17, female)

One of five mystery shopper had no difficulty in finding the products on her list, even the ones that were not on her list before due to the swap:

- “I found the products rather quick” (resp. #6, female)

Two of five mystery shoppers questioned the tastiness of the swapped target product, even though in their original checklist ‘*Ontbijtkoek Gember*’ was never mentioned:

- “The ‘Ontbijtkoek Gember’ [swapped target product] sounds terrible, do people actually eat that?” (resp. #10, male)
- “I did not know they sold ‘Ontbijtkoek Gember’ [swapped target product], it doesn't sound tasty” (resp. # 14, female)

One of two mystery shoppers who spoke up to the research leader was invited to participate in the follow up interview. He noticed the swap:

- “There are different products on my list then before, did you switch them? That was pretty obvious” (resp. #19, male)

He also gave two possible explanations for the occurrence of incorrect reports:

- “It's possible to take the easy way out. You (research leader) can't check what I (mystery shopper) did or didn't see” (resp. #19, male)
- “It could be ‘scary’ for a freshman to speak up to the research leader” (resp. #19, male)

True Checklist / No Swap (control group)

In total, 17 mystery shoppers were not confronted with any form of manipulation. Of the seventeen, 17.6% (*n* = 3) was invited to participate in the follow up interview. In total, 11.7% (*n* = 2) took initiative and spoke up to the research leader about perceived difficulties. This leaves 88.2% (*n* = 15) of the mystery shoppers who didn't initiate any form of interaction with the research leader.

All three mystery shoppers who were invited to participate in the follow up interview stated that even though the experiment seemed difficult it was doable:

- "If you don't take it seriously and just quickly read through the checklist just to get it over with, than I understand it's difficult. If you take your time it is pretty easy" (resp. #5, female)
- "It seemed more difficult than it really was" (resp. #9,male)
- "Some people are better in remembering than others. On the other hand, I'm not very good in remembering in general and I could do it. I think it depends how serious you take this" (resp. #13, male)

One of three mystery shoppers who were invited to participate gave a possible explanation for the occurrence of incorrect reports among mystery shoppers. He stated that the lack of motivation could be a possible explanation:

- "If you only do it for the reward (the study credits), you could just fill something out and leave." (resp. #13, male)

One of two mystery shoppers who spoke up to the research leader was invited to participate in the follow up interview. He agreed that the experiment was difficult, but doable. Moreover, he stated that two filler products were difficult to find.

- "It was difficult to remember all the products, but if you took your time it was doable" (resp. #1, male)
- "The Campina [Filler product D] and Chocomel [Filler product E] were difficult to find" (resp. #1, male)

DISCUSSION

During the past decades, researchers have attempted to identify the reliability of the mystery shopping method; a facet which is essential is the reliability of mystery shopping reports. These reports are often based on human memory, or researcher cognition. The objective of this study was to identify to what extend two forms of researcher cognition bias (*disconfirmed expectancies* and *exposure to misinformation*) influenced the reliability of mystery shopping reports.

Disconfirmed expectancies suggests that mystery shoppers who's expectancies are disconfirmed must either discard the disconfirmed expectancy or justify why it has not been disconfirmed ("I believe [X], but I observed [Y]").

Exposure to misinformation suggests that mystery shoppers can be misled by post-event suggestions following an observed event ("I remembered [X], but I was supposed to remember [Y]").

Disconfirmed Expectancies - Exposure to Misinformation (experimental group)

Sixteen mystery shoppers were confronted with both forms of researcher cognition bias: disconfirmed expectancies in the form of a false checklist before the visit and exposure to misinformation in the form of the swap of the checklist during their visit. According to the three phases of memory of Morrison et al. (1997) these mystery shoppers were both manipulated in the storage and retrieval phase of memory. In contrary to expectations, the number of correct reports did not significantly differed from other experimental groups in which only one manipulation occurred (H1). One might expect that mystery shoppers who were confronted with two forms of researcher cognition bias would have more difficulty reporting correctly in comparison to mystery shoppers who were confronted with one form of researcher cognition bias. This condition was confronted with psychological discomfort in both the storage and retrieval phase of memory from manipulations and thus had two phases where they could report incorrectly. The fact that this combination of biases did not further diminished the number of correct reports could be considered as somewhat favorable for the reliability of mystery shopping reports as apparently the rise in number of researcher cognition biases does not further diminishes the number of correct reports.

Furthermore, this was the only experimental group were mystery shoppers did not provide (un)intentionally false information during the follow up interviews.

The mystery shoppers in this condition performed significantly less correct reports for the target product in comparison to filler products [A] and [C], but they did not perform significantly less correct reports for the target product in comparison to filler products [B] and [D]. These findings suggest that a form of halo effect occurred in which manipulation effects emitted to surrounding parts of the checklist, the part of the checklist that was actually available in store. The halo-effect is a cognitive bias in which global evaluations bleed over into judgments about specific traits (Nisbett & Wilson, 1977). The mystery shoppers could have had a negative judgment of the target product item, due to psychological discomfort created by disconfirmed expectancy: the absence of the target product. This could have led to a negative judgment of surrounding items on the checklist, in this case filler product [B] and [D]. Due to exposure to misinformation mystery shoppers had to recreate the event after their visit; this condition showed that effects of manipulations emitted to surrounding items.

Results of other experimental groups showed that the halo-effect for filler product [B] derived from disconfirmed expectancies, as exposure to misinformation only affects the target product. However, the combination between these forms of bias resulted in a significant drop for the number of correct reports for filler product [D].

Five mystery shoppers spoke up to the research leader about experienced difficulties during their visit. Mystery shoppers only discussed experienced difficulties from exposure to misinformation bias and did not mention their disconfirmed expectancies. Results from other experimental groups suggest that exposure to misinformation ‘overruled’ effects of disconfirmed expectancies for reasons to speak up, as exposure to misinformation occurred in the retrieval phase which is one phase ‘later’ than the storage phase of memory. Thus, the mystery shoppers ‘forgot’ effects from disconfirmed expectancies.

Disconfirmed Expectancies - No Exposure to Misinformation (experimental group)

Results supported the hypothesis that fifteen mystery shoppers who were disconfirmed in their expectancies performed significantly less correct reports than mystery shoppers who were not disconfirmed in their expectancies (H2). Mystery shoppers performed significantly less correct reports for the target product and filler product [B] in comparison to other filler products. These findings suggested that a form of halo effect (Nisbett & Wilson, 1977) occurred in which effects of disconfirmed expectancies emitted to a filler item above, a product item that was actually available in store. These mystery shoppers were only manipulated in the storage phase, Morrison et al. (1997) stated that memories in the storage phase could interfere with competing memories. The effect of disconfirmed expectancy could have resulted in distortions during the retrieval phase of memory which resulted in a significant drop of correct reports for the above filler item [B].

Even though the target product was nowhere to be found in store 53.3% (n = 8) of mystery shoppers managed to note a price for the target product and did not speak up about the unfindable product to the research leader. According to Hasher and Greenberg (1977) mystery shoppers could have dealt with their psychological discomfort through the justification that they had not been disconfirmed. In other words, they possibly convinced themselves that they had seen the product even though they it did not exist (Festinger, 1985). On the other hand they could possibly express some form of social desirable answering by meeting the demands of the researcher. The conformity experiment of Asch (1955) that demonstrated the degree to which an individual's own opinions is influenced by an expert even though the opinion of the expert is obviously wrong. In this example the mystery shoppers could have wanted to meet the demands of the researcher even though this contradicted their own observation. Either way, the fact that they (un-)intentionally gave false information about the 'findability' of a non-existing product could question the reliability of these mystery shoppers report:

- "It was quite a search but eventually I found them all" (resp. #3, female)

The remaining 46.7% (n = 7) of the mystery shoppers who were disconfirmed in their expectancies spoke up to the research leader and dealt with their psychological discomfort by discarding their disconfirmed expectancy.

- "I could not find the Taksi [target product], does that product even exist?" (resp. #16, male)

Even though numbers of mystery shoppers speaking up to the research leader did not significantly differed between the conditions, there was a trend visible in which mystery shoppers in this condition might speak up more to research leaders in comparison to other conditions. An explanation for this trend might be that it could be considered less confrontational, or 'scary', to speak up to a research leader about experienced difficulties from external factors (e.g. research material) than to speak up about experienced difficulties from internal factors (e.g. own failure).

Results showed that mystery shoppers who were disconfirmed in their expectancies took more time to find products in store, which is not unexpected as the target product did not exist. This difference could have occurred because some mystery shoppers already were familiar with the specific store of this study. However, Turley and Milliman (2000) stated that shopping atmospherics are designed by 'general rules', especially supermarkets (vegetable section in the front; bread section in the back, etc.). Therefore one could argue that all mystery shoppers were familiar with the atmospherics of the store and thus explaining the difference in time to find products as a result of the manipulation. One might reason that based on these findings one can predict whether mystery shoppers experienced difficulties during their visit based on the time it took them to perform the visit.

No Disconfirmed Expectancies - Exposure to Misinformation (experimental group)

Results of this study supported the hypothesis that the fifteen mystery shoppers who were exposed to misinformation performed significant less correct reports than mystery shoppers who were not exposed to misinformation (H3). These mystery shoppers were only manipulated in the retrieval phase of memory (Morrison et. al, 1997).

Moreover, some criticism can be stated on the way mystery shoppers dealt with exposure to misinformation. Results of this study showed that only 13.3% (*n* = 2) of the mystery shoppers spoke up to the research leader about experienced difficulties (e.g. swap of checklist).

- "There are different products on my list than before, did you swap them?" (resp. #19, male)

This leaves 86.7% (*n* = 13) of mystery shoppers who did not speak up to the research leader about the different version of the checklist and filled out the target product item even though they could have never seen that product during their visit. These findings are in line with expectations that exposure to misinformation can lead people to recall seeing objects that did not appear or occur in the original event (Loftus & Pickrell, 1995; Nourkova, Bernstein, & Loftus, 2004). Two mystery shoppers possibly confirmed existence of a product that they had never seen. One mystery shopper confirmed that she found all the products 'rather quick', thereby possibly confirming the existence of the target products. However, these comments could also be seen as forms of social desirable answering for meeting the researcher's demands as they might not want to express experienced difficulties based on internal factors (e.g. own failure) (Asch, 1955). Nonetheless, these comments raised some concerns about the reliability of mystery shopping reports as they (un)intentionally provided the research leader with false information.

- “The ‘Ontbijtkoek Gember’ [swapped target product] sounds terrible, do people actually eat that?” (resp. #10, male) & (resp. # 14, female)
- “I found the products rather quick” (resp. #6, female)

Thus, there were thirteen mystery shoppers that didn’t speak up to the research leader about the different version of the checklist after their visit. One might argue that these mystery shoppers did not notice different product items on their checklist after their visit. However, the statistical analyses showed that mystery shoppers who were exposed to misinformation took significant more time to fill out the checklist after their visit. This difference could occur from dealing with psychological discomfort from the manipulation. This is in line with the findings of Ayers and Reder (1998) who suggested that response time of humans who are exposed to misinformation significantly drops. One might reason that based on these findings one can predict whether mystery shoppers experienced difficulties during after visit based on the time it took them to fill out the checklist.

The findings of a form of halo-effect, as seen by disconfirmed expectancies, did not occur. These findings were not in line with expectations derived from the halo effect theory (Nisbett & Wilson, 1977), which suggests that the negative judgment of an item can bleed over into other items. Mystery shoppers performed significantly less correct reports for the target product, but performed statistically equal for filler products. In other words, negative effects of the swap of the checklist did not emit to surrounding items. This could be explained by the fact that the effect of manipulation occurred in the retrieval phase of memory, thus after the visit instead of during the visit. The correct information for filler products could have already been correctly storaged during the storage phase of memory (Morrison et al., 1997). Apparently effects of exposure to misinformation in the retrieval phase did not influence, or ‘overrule’, the storaged memories. As no halo-effect was found in this condition, the halo-effect can be ascribed to the effect of disconfirmed expectancies.

No Disconfirmed Expectancies - No Exposure to Misinformation (control group)

This study supported the hypothesis that seventeen mystery shoppers who were not confronted with any form of researcher cognition bias performed significantly more correct reports than mystery shoppers who were confronted with a form of researcher cognition (H4). With an average of 4.24 correct reports out of 5 (85% correct report rate), these findings suggest that as long as expectancies of mystery shoppers are confirmed and mystery shoppers are not exposed to misinformation they can provide an exact accurate reporting rate of 85% of their findings. However, they were not flawless. Miller (1956) proposed that seven, plus or minus two items, is the limit within the human mind. In this study mystery shoppers had to remember five prices, each price consisting of three numbers (e.g. €1,39). Therefore one could argue that they had to remember fifteen numbers, which makes an 85% correct report rate acceptable to standards of Miller (1956). Moreover, mystery shoppers had to remember a checklist that contained thirty-two items in total. According to Miller’s standards the tasks of these mystery shoppers could be labeled as difficult, as the amount of items that had to be remembered was almost fivefold to what Miller proposed to be the limit within the human mind. Although it was a difficult experiment to participate in, the results of the experiment and the follow up interview suggested that unbiased mystery shoppers had little difficulty to provide an 85% accurate report rate.

- “It seemed more difficult than it really was” (resp. #9, male)

Speak up or remain silent

An important part of this study was that mystery shoppers were given the opportunity to speak up to the research leader about experienced difficulties during or after the visit. Speaking up about experienced difficulties would overrule an incorrect report and therefore would result in a rise of correct reports.

In total, sixteen mystery (25.4%) shoppers spoke up to the research leader about experienced difficulties. Speaking up in this study could be seen as a form of employee voice. Employee voice refers to the situation where employees speak up to express concerns, opinions, or suggestions about their own work situation or organization (Van Dyne, Ang, & Botero, 2003). Most of the mystery shoppers in this study who spoke up wanted to express concerns and make sure that the research leader understood that their experienced difficulties were all based on external factors (e.g. problems with the material).

- "Filler product [D] and [E] were difficult to find" (resp. #1, male)

Mystery shoppers who spoke up to the research leader also provided insights into why a mystery shopper would not speak up. With these comments, in a way, they wanted to 'help' the research leader by offering their concerns to improve future data collection methods.

- "It could be 'scary' for a freshman to speak up to the research leader" (resp. #19, male)

In total, 47 mystery shoppers (74.6%) remained silent when confronted with researcher cognition biases, which can be seen as a form of employee silence. Employee silence refers to situations where employees (intentionally) withhold information that might be useful to the organization, which can happen if employees do not speak up to a supervisor (Milliken & Morrison, 2003). According to Brinsfield (2013) employees can have different motives to remain silent: *disengaged employee silence* and *diffident employee silence*.

Disengaged silence refers to the lack of interest and motivation to speak up (Brinsfield, 2013). The exploration of motivations among mystery shoppers for participating in mystery shopping visits; both intrinsic (participating for the experience) and extrinsic (participating for the reward) can be a useful to better understand why mystery shoppers perform in certain ways (Allison, 2009; Allison, Severt, & Dickson, 2010). Examples of disengaged silence are: "The issue did not personally affect me"... "I did not care what happened". Two mystery shoppers confirmed the assumptions that a lack of motivation or interest could lead to an increase in incorrect reports.

- "If you only do it for the reward (the credits), you could just fill something out and leave" (resp. #13, male).
- "If you don't take it seriously and just quickly read through the checklist just to get it over with, than I understand it's difficult" (resp. #5, female)

Both comments show that a lack of intrinsic motivation (participating for the experience) in combination with a form of disengaged silence ('this does not affect me personally') could result in a rise of incorrect reporting among mystery shoppers.

Diffident silence refers to the hesitance to speak up through a lack of self-confidence (Brinsfield, 2013). Examples of diffident silence are: "I did not feel confident enough to speak up"…"I felt insecure"…"I wanted to avoid embarrassing myself". Mystery shoppers who did not spoke up to the research leader admitted during follow up interviews that they encountered some difficulties, in which they 'blamed' their self. This silence could be seen as a form of diffident silence:

- "When I returned I noticed that I remembered the wrong products. I was really starting to doubt myself" (resp. #4, male)
- "I didn't want to disturb you (the researcher) with an error that I made" (resp. #15, male)

Comments of mystery shoppers showed that a lack of confidence in their own observations prevented them from speaking up to the research leader about their experience difficulties. Noelle-Neumann (1974) suggested that feelings of self-doubt may discourage people from expressing ideas that fail to conform to public opinions. In other words, due to a lack of confidence in their own observations these mystery shoppers chose to conform to the demands of the research leader. In this example the mystery shoppers wanted to meet the demands of the researcher even though their own observation contradicted the presented items on the checklist. These forms of diffident silence could result in a rise of incorrect report among mystery shoppers who are confronted with difficulties, as long as the “threshold” to break with the social conformity remains too high.

General findings

Measurements to which mystery shoppers could score a ‘correct report’ was narrow, only exact correct prices were considered to be correct reports. Reason for this was to create a clear baseline assessment for the number of correct reports mystery shoppers could perform. Allowing ‘one number to be off’ (e.g. €1,40 instead of €1,39) would enhance ‘educated-guess-bias’, as the sample was expert concerning the products and the location of visit (supermarket). Goettler et al. (2009) investigated that experts guessed significantly better than chance prediction. As the sample was experienced users of supermarkets (students) this educated guess bias was limited to such an extent that it could not interfere for the baseline assessment, therefore the price had to be exactly correct.

In general, one may question the reliability of mystery shopping reports as mystery shoppers showed various reasons (lack of motivation and lack of self-confidence in owns observation) for not speaking up about experienced difficulties. As long as “thresholds” for speaking up remain too high, the number of incorrect reports will probably not diminish. On the other hand, completely accurate mystery shopping materials (e.g. checklist) diminished motives for speaking up. A balance has to be found between accuracy of checklists and levels of speaking up about experienced difficulties.

Limitations

Limitations for this study include the excluding of a target product item and various common method biases.

Originally the checklist contained six product items (two target products and four filler products). One target product (Peijenburg Ontbijtkoek Naturel) showed a significantly lower number of correct reports. Reasons for this significant drop of correct reports could arose from a mismatch between product name on checklist and product name in store; and four similar products for 'Peijenburg Ontbijtkoek Naturel'. This resulted in a wide variety of product reports (21 different prices), with a significantly higher standard deviation. Based on these findings the target product item was excluded.

Common method biases contributed to variance due to measurement method instead of measurement construct (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). All potential sources of bias were considered and were minimized where possible (see Table 15).

Table 15. ***Potential method biases, with the source of bias, minimization effort and remaining limitation***

Method bias*	**Source of bias***	**Minimization effort**	**Remaining limitation**
Consistency Effect	Respondent attempts to maintain consistency of responses.	Mystery shoppers could not view questions/items from separate sections simultaneously (trial visit, actual visit, follow up interview)	Mystery shoppers had the ability to alter previous answers.
Social Desirability	Participants attempt to provide answers that are socially acceptable or what is perceived the researcher wants.	The researcher leader refrained from affirming or disputing comments made by the mystery shopper.	With the measurement of experienced difficulties, social desirability bias cannot be completely eliminated.
Participant Mood	Participants' mood at the time of the study.	Mystery shoppers were allowed to choose the time and day for the experiment to avoid unnecessary strain.	Mystery shoppers were not asked about their mood (sleepy, active, etc.). This could have influenced results.
Item Complexity	Items in the measurement instrument with unclear meanings.	The checklist was predetermined, rehearsed and pretested	One target product item deemed to 'unclear' to properly report (Peijenburg Ontbijtkoek).
Item Priming Effect	Introduction of an item in the checklist could alter completion of similar, subsequent items.	Mystery shoppers were provided with all the information regarding the experiment prior to the visit about the types of items being studied.	Due to the manipulation of the swapped checklist the item priming effect remained.

***** As identified by Podsakoff, MacKenzie, Lee, & Podsakoff (2003)

Future research

The majority of the sample consisted of *novice* mystery shoppers according to the classification scheme of Allison (2009) because all, except four, had performed less than ten mystery shop visits. The four who had could be seen as *exploratory* mystery shoppers. *Career*, or professional, mystery shoppers who performed more than 25 visits, did not participate in this study. For future research it might be interesting to study the effect of the forms of researcher cognition bias on mystery shopper professionals. Mystery shopper professionals should be trained in memory coding techniques and therefore should perform more correct reports according to Miller (1956). Also, professional mystery shoppers might have other motivations for participating in mystery shopping studies.

Follow up interviews showed the importance of '*breaking the silence*' among mystery shoppers. Research exploring relationships between mystery shoppers and research leaders might provide new insight into the motivations to (not) speak up about experienced difficulties during visits. Additionally, research exploring the relationship between the level of intrinsic motivation and the willingness to speak up could contribute to a better understanding and selecting of mystery shoppers for future studies.

In this study, mystery shoppers were primed on the trustworthiness of materials and research leader. They were not aware that possible difficulties could arise. One might argue that a 'warned' mystery shopper would report difficulties more often. For future research it might be interesting to study the effect of different forms of instruction before mystery shop visits on speaking up or remaining silent among mystery shoppers.

Practical implications

The greatest factor threatening reliability of mystery shopping reports is employee silence. In general, mystery shoppers who experience difficulties during or after their visits will not speak up. Therefore, the mystery shopping industry should take initiative and always ask mystery shoppers about difficulties during or after mystery shopping visits. Something as simple as; *"How did it go?"* or *"Did you experience any difficulties?"* proved useful to discover experienced difficulties, which could otherwise stay hidden.

The mystery shopping industry can use results provided in this study to create better tailored checklists for mystery shopping visits. Mystery shopping firms should evaluate their current checklists, weighing perceived costs but keeping in mind that an inaccurate checklist could result in a rise of inaccurate reports.

This study showed that there were significant differences in time it took to complete the visit and time it took to complete the checklist for mystery shoppers who were confronted with forms of researcher cognition bias. The mystery shopping industry should include timing in mystery shopping materials. A significant rise in time to complete visits or checklists could indicate that mystery shoppers experienced difficulties.

CONCLUSION

This study demonstrated that multiple forms of researcher cognition bias had a significant negative effect on the reliability of mystery shopping reports:

- The confrontation with multiple forms of researcher cognition bias resulted in a significant drop of correct reports, but did not result in a drop of incorrect reporting in comparison to mystery shopper who were only confronted with one form of researcher cognition bias;
- When mystery shoppers are confronted with disconfirmed expectancies the number of correct reports significantly drops. Moreover, they show a form of halo-effect; not only does the number of correct reports drop significantly for ‘non-existing’ items, but also for surrounding ‘existing’ items;
- When mystery shoppers are exposed to misinformation the number of correct reports will only drop for ‘non-existing’ items, not for the ‘existing’ items;
- Mystery shoppers who were not confronted with any form of researcher cognition bias showed an exact correct reporting rate of 85% for narrow measurement items;
- High times to complete a mystery shop visit and checklist could indicate for experienced difficulties among mystery shoppers;
- Mystery shoppers ‘speak up’ to express concerns about external factors (e.g. inaccurate materials), rather than internal factors (e.g. own failure);
- Mystery shoppers showed forms of intentionally employee silence originating from lack of intrinsic motivation and lack of self-confidence in owns observation results in social conformity.

In conclusion, the mystery shopping method can be considered as liable for incorrect reporting as long as “*thresholds*” to break with mystery shopper silence remain too high.

LITERATURE

Allison, P. B. (2009). *Mystery Shopping Motivations and the Presence of Motivation Crowding.* Florida: Department of Educational Research, Technology and Leadership at the University of Central Florida.

Allison, P. B., Severt, D., & Dickson, D. (2010). A Conceptual Model for Mystery Shopping Motivations. *Journal of Hospitality Marketing & Management, 19*(6), 629-657.

Anderson, D. N., Groves, D. L., Lengfelder, J., & Timothy, D. (2001). A research approach to training: a case study of mystery guest methodology. *International Journal of Contemporay Hospitality Management, 13*(2), 93-102.

Asch, S. E. (1955). Opinions and Social Pressure. *Scientific America, 193*(5), 31-35.

Mystery Shopper Providers Association. (2013, October). *http://www.mysteryshop.org*.

Ayers, M. S., & Reder, L. M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review, 5*(1), 1-21.

Baker, C. H. (1961). Maintaining the level of vigilance by means of knowledge of results about a secondary vigilance task. *Ergonomics, 4*, 311-316.

Berkowitz, S. R., Laney, C., Morris, E. K., Garry, M., & Loftus, E. F. (2008). Pluto Behaving Badly: False Beliefs and Their Consequences. *The American Journal of Psychology, 121*(4), 643-660.

Boon, J. C., & Davies, G. M. (1993). The influence of biographical information on event memory: When it’s not what you know but whom you know. *Journal of General Psychology, 120*, 517-530.

Brinsfield, C. T. (2013). Employee silence motives: Investigation of dimensionality and development of measures. *Journal of Organizational Behavior, 34*, 671-697.

Carlsmith, J. M., & Aronson, E. (1963). Some Hedonic Consequences of the Conformation and Disconformation of Expectancies. *Journal of Abnormal and Social Psychology, 66*(2), 151-156.

Dawes, J., & Sharp, B. (2000). The Reliability and Validity of Objective Measures of Customer Service; Mystery shopping. *Australian Journal of Market Research*, 1-23.

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

Festinger, L. (1957). *A Theory of Cognitive Dissonance.* Stanford: Stanford University Press.

Finn, A. (2001). Mystery shopper benchmarking of durable-goods chains and stores. *Journal of Service Research, 3*(4), 310.

Finn, A., & Kayande, U. (1999). Unmasking a phantom; A psychometric assessment of mystery shopping. *Journal of Retailing, 75*(2), 195-217.

Ford, R. C., Latham, G. P., & Lennox, G. (2011). A new tool for coaching employee performance improvement. *Organizational Dynamics, 40*, 157-164.

Goettler, C. E., Waibel, B. H., Goodwin, J., Watkins, F., Toschlog, E. A., Sagraves, S. G., et al. (2010). Trauma Intensive Care Unit Survival: How Good Is an Educated Guess. *The Journal of Trauma: Injury, Infection, and Critical Care, 68*(6), 1279-1288.

Gosselt, J. F., van Hoof, J. J., de Jong, M. D., & Prinsen, S. (2007). Mystery Shopping and Alcohol Sales: Do Supermarkets and Liquor Stores Sell Alchol to Underage Customers. *Journal of Adolescent Health, 41*, 302-308.

Hasher, L., & Greenberg, M. (1977). Expectancies as a Determinant of Interference Phenomena. *The American Journal of Psychology, 90*(4), 599-607.

Heaps, C. M., & Nash, M. (2001). Comparing Recollective Experience in True and False Autobiographical Memories. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*(4), 920-930.

Hesselink, M., & van der Wiele, T. (2003, March). Mystery Shopping: In-depth measurement of customer satisfaction. Erasmus Research Institute of Management (ERIM).

Holmes, D. S. (1972). Effects of Grades and Disconfirmed Grade Expectancies on Students' Evaluations of their Instructor. *Journal of Educational Psychology, 63*(2), 180-183.

Kocevar-Weidinger, E., Benjes-Small, C., Ackermann, E., & Kinman, V. R. (2009). Why and how to mystery shop your reference desk. *Reference Services Review, 38*(1), 28-43.

Leippe, M. R., Eisenstadt, D., Rauch, S. M., & Seib, H. M. (2004). Timing of Eyewitness Expert Testimony, Jurors' Need for Cognition, and Case Strength as Determinants of Trial Verdicts. *Journal of Applied Psychology, 89*(3), 524-541.

Leippe, M. R., Eisenstadt, D., Rauch, S. M., & Stambush, M. A. (2006). Effects of Social-Comparative Memory Feedback on Eyewitnesses' Identification Confidence,Suggestibility, and Retrospective Memory Reports. *Basic and Applied Social Psychology, 28*(3), 201-220.

Lindsay, D. S. (1990). Misleading Suggestions Can Impair Eyewitnesses' Ability to Remember Event Details. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*(6), 1077-1083.

Lindsay, D. S., & Johnson, M. K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition, 17*(3), 349-358.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year inves tigation of the malleability of memory. *Learning & Memory, 12*, 361-366.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic Integration of Verbal Information into a Visual Memory. *Journal of Experimental Psychology: Human, Learning and Memory, 4*(1), 19-31.

Macrea, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology, 23*, 77-87.

Miller, G. A. (1956). The magic number seven plus or minus two: Some limits on our capacity to process information. *Psychological Review, 63*, 81-97.

Milliken, F. J., & Morrison, E. W. (2003). Shades of Silence: Emerging Themes and Future Directions for Research on Silence in Organizations. *Journal of Management Studies, 40*(6), 1546-1568.

Morgan, C. A., Southwick, S., Steffian, G., Hazlett, G. A., & Loftus, E. F. (2013). Misinformation can influence memory for recently experienced, highly stressful events. *International Journal of Law and Psychiatry, 36*, 11-17.

Moriarty, H., McLeod, D., & Dowell, A. (2003). Mystery shopping in health service evaluation. *British Journal of General Practice, 53*, 942-946.

Morrison, L. J., Colman, A. M., & Preston, C. C. (1997). Mystery customer research: Cognitive processes affecting accuracy. *Journal of Market Research Society, 39*(2), 349-360.

Ng Kwet Shing, M., & Spence, L. J. (2002). Investigating the limits of competitive intelligence gathering: Is mystery shopping ethical? *Business Ethics: A European Review, 11*(4), 343-353.

Nisbett, R. E., & Wilson, T. D. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review, 84*(3), 231-259.

Noelle-Neumann, E. (1974). The Spiral of Silence A Theory of Public Opinion. *Journal of Communication, 24*(2), 43-51.

Nourkova, V., Bernstein, D. M., & Loftus, E. F. (2004). Altering traumatic memory. *Cognition and Emotion, 18*(4), 575-585.

Pezdek, K., Finger, K., & Hodge, D. (1997). Planting false childhood memories: The role of event plausibility. *Psychological Science, 8*, 437-441.

Pezdek, K., Lam, S. T., & Sperry, K. (2009). Forced confabulation more strongly influences event memory if suggestions are other-generated than self-generated. *Legal and Criminological Psychology, 14,* 241-252.

Pezdek, K., Sperry, K., & Owens, S. (2007). Interviewing witnesses: The effect of forced confabulation on event memory. *Law & Human Behavior, 31*, 463-478.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903.

Schreiber, T. A., & Sergent, S. D. (1998). The role of commitment in producing misinformation effects in eyewitness memory. *Psychonomic Bulletin & Review, 5*(3), 443-448.

Struyk, R. J., & Haddaway, S. R. (2011). Licensed Lenders' Services to Indonesian SMEs: 'Mystery Shopping' Results. *International Journal of Economics and Finance, 3*(3), 3-14.

Tarantola, C., Vicard, P., & Ntzoufras, I. (2012). Monitoring and improving Greek banking services using Bayesian Networks: An analysis of mystery shopping data. *Expert Systems with Applications, 39*, 10103-10111.

Taylor, D. A., Altman, I., & Sorrentino, R. (1969). Interpersonal Exchange as a Function of Rewards and Costs and Situational Factors: Expectancy Confirmation-Disconfrmation. *Journal of Experimental Social Psychology, 5,* 324-339.

Turley, L. W., & Milliman, R. E. (2000). Atmospheric Effects on Shopping Behavior: A Review of the Experimental Evidence. *Journal of Business Research, 49*, 193-211.

Van der wiele, T., Hesselink, M., & Van iwaarden, J. (2005). Mystery Shopping: A Tool to Develop Insight into Customer Service Provision. *Total Quality Management, 16*(4), 529-541.

Van Dyne, L., Ang, S., & Botero, I. C. (2003). Conceptualizing Employee Silence and Employee Voice as Multidimensional Constructs. *Journal of Management Studies, 40*(6), 1360-1392.

Wilson, A. M. (1998). The use of mystery shopping in the measurement of service delivery. *The Services Industries Journal, 18*(3), 148-163.

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

APPENDICES

APPENDIX I: Interviews (N = 19)

The interviews are sorted per condition and only available in Dutch:

Disconfirmed Expectancies (*n* = 4 | 3 interview + 1 spoke up)

3. Respondent # (non-existing – no swap) - Female

O En hoe vond je dat het ging?
R Ja ging goed.
O Heb je alles makkelijk kunnen vinden?
R Ja ik ben wel lang weggeweest of niet? Het was even zoeken, maar heb uiteindelijk alles gevonden. Die prijzen onthouden was wel pittig. Ik ben wel drie keer teruggelopen bij sommige producten.
O Weet je nog bij welke producten dat was?
R Ik kreeg de campina vlavlip maar niet in m'n hoofd. Dat personeel moet wel gedacht hebben die weet ook niet wat ie wil.
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Valt het niet op met die tas? Er komen nu een aantal mysteryshoppers per dag die winkel binnen met dezelfde tas die allemaal een hamburger bestellen, dat valt op lijkt me.

7. Respondent # (non-existing – no swap) – Male

O En hoe vond je dat het ging?
R Ja ging goed, het leek moeilijker dan dat het was. Vond het eerste gedeelte moeilijker met die lange vragenlijst na afloop. Ik weet echt niet alles zeker meer. Is dat erg?
O Daar ging het meer om de indruk die de winkel achterliet. Hoe ging het tweede gedeelte, weet je daar wel alles zeker?
R Ja, heb alle zes de producten snel gevonden en dan was het alleen nog een kwestie van die prijzen stampen.
O Ben je nog tegen iets van moeilijkheden aangelopen?
R Nee
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Valt het niet op dat er zoveel mysteryshoppers daar een bezoek doen? Volgens mij had de slagerij mij door namelijk.

11. Respondent # (non-existing – no swap) - Female

O En hoe vond je dat het ging?
R Wel goed denk ik. Kon alleen twee producten niet vinden.
O Hoe dat zo?
R De Peijenburg en Taksi waren er niet, of stonden op een compleet andere plek. Heb het hele schap doorgezocht maar kon ze niet vinden. Dus of de producten zijn afwezig of ik heb echt verkeerd gezocht.
O Goed dat je het zegt, ik zal er straks even naar gaan kijken. De producten moeten namelijk wel aanwezig zijn.
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Buiten die producten om was er niks aan de hand. Ging allemaal soepel. Die tas is trouwens wel echt lelijk. Kan eigenlijk echt niet voor jongens.

16. Respondent (non-existing – no swap) – Spoke Up - Male

R Ik kon de Taksi niet vinden. Bestaat die wel?
O Maar je hebt de rest van de producten wel gevonden?
R Ja die wel. Campina was nog even zoeken, maar uiteindelijk kwam ik er achter dat er een apart schap met vla was. Die had ik in eerste instantie niet gezien. Vond het eerste gedeelte wel leuker trouwens.
O Hoe dat zo?
R Daar mocht je tenminste praten met het personeel. Vond het in tweede gedeelte apart dat je als klant lang op zoek bent naar producten maar dan niet tegen het personeel mag praten. Een medewerker vroeg mij of

hij mij kon helpen omdat ie denk ik door had dat ik iets niet kon vinden. Toen moest ik dus nee zeggen, terwijl ik daarna nog dik vijf minuten door die winkel heb lopen slenteren. Heb toen besloten de taksi te laten voor wat het was.
O	Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R	Die tas kan echt niet! En zou het personeel echt niks door hebben? Ik betwijfel het.

Exposure to Misinformation (*n* = 6 | 5 interviews + 1 spoke up)

2. Respondent (existing – swap) - Female

O	En hoe vond je dat het ging?
R	Was echt wel moeilijk, hoop dat ik alles goed heb onthouden.
O	Ben je bang dat je sommige producten niet goed hebt?
R	Ja, was best wel raar. Had bij twee producten de verkeerde onthouden kennelijk. Ik had tropisch fruit ipv bosvruchten en naturel ontbijtkoek ipv gember. En toen kwam ik terug stonden er opeens andere producten. Maar heb het op t formulier er bij gezet. Het is nog vroeg he. Hoop dat t niet erg is?
O	Nee, goed dat je het zegt.
R	Voor de rest is alles volgens mij goed gegaan.
O	Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R	Die vrouw achter het vlees (eerste deel onderzoek) was wel heel aardig, wellicht dat ze mij door had.

6. Respondent (existing – swap) - Female

O	En hoe vond je dat het ging?
R	Ja, ging wel goed. Leek in het begin wel moeilijk toen ik al die producten zag. Ook met al die prijzen. Maar volgens mij heb ik het wel goed gedaan.
O	Heb je alles makkelijk kunnen vinden?
R	Ja, was zo gevonden. Heb vroeger in supermarkt gewerkt dus ik wist waar alles ongeveer zou liggen.
O	Het blijkt inderdaad dat je bezoektijd kort was, maar de invultijd voor de lijst was dan weer lang. Enig idee waarom?
R	Heb ik er zo lang over gedaan ja? Had ik zelf niet door. Heb er geen speciale verklaring voor.
O	Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R	Nee.

10. Respondent (existing – swap) - Male

O	En hoe vond je dat het ging?
R	Het ging goed. Het personeel was heel aardig. De kassière hielp mij heel netjes en het was niet druk in de winkel.
O	En hoe ging het onderzoek zelf?
R	Goed, die hamburger halen was wel wat raar. Ze wist niet eens hoe ze de hamburger moest bakken, die kunnen ze beter ontslaan.
O	En het tweede gedeelte van het onderzoek?
R	Dat was moeilijker zoals je al aangaf, maar nog steeds te doen. Ik dacht wel dat de kassière van de andere kassa mij door had. Ze zei iets van ‘daar heb je d’r weer één’. Maar ik heb net gedaan of ik het niet hoorde.
O	En heb je alle producten makkelijk kunnen vinden?
R	Ja, de chocomel was wel op. Maar het prijskaartje was goed zichtbaar.
O	Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R	Die ontbijtkoek gember lijkt me echt niet te eten, zijn er echt mensen die dat kopen?

14. Respondent (existing – swap) - Female

O	En hoe vond je dat het ging?
R	Wel redelijk. Vond het toch wel lastig om al die prijzen en producten te onthouden. Het scheelde wel dat ik met het eerste onderzoek een keer heb kunnen oefenen. Daarmee was de spanning voor de tweede keer er wel af.
O	Heb je alles kunnen vinden?
R	Ja, het was wel even zoeken hoor. Vooral Peijenburg was lastig te vinden.
O	Koop je normaal geen ontbijtkoek gember?
R	Nee, wist trouwens ook niet dat ze dat daar verkochten. Lijkt me niet lekker.

O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Er stonden net jongeren voor de deur (op het parkeerterrein), die maakte veel lawaai daardoor kon ik me moeilijk concentreren.

17. Respondent (existing – swap) - Female

O En hoe vond je dat het ging?
R Prima, leuk onderzoek.
O Zou je iets kunnen uitweiden?
R Het is weer eens wat anders. Ik heb tot nu toe alleen maar van die vragenlijsten onderzoeken gedaan en dit is mooi praktisch.
O En hoe ging het onderzoek doen jou af?
R Het was het eerste bezoek wel even spannend. Je hebt de hele tijd het idee dat iedereen je door heeft. Dat is natuurlijk niet zo, maar dat lijkt zo. Het gesprek met de slagerij kwam op mij wel een beetje fake over. Het leek alsof zij het door had.
O En het tweede gedeelte?
R Was voor mijn gevoel natuurlijker, omdat je nu geen fake gesprek hoefde aan te gaan met het personeel. Maar omdat het lastig om alles te onthouden moet je een aantal keer teruglopen. Dit moet het personeel door gehad hebben. Het is toch raar als je een kwartier rondloopt en dan een zak chips koopt. Ook met die tas om moet het duidelijk zijn geweest dat ik geen echte klant was.
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Ik zou het onderzoek doen bij meerdere filialen en niet steeds bij dezelfde. Dan valt het minder op. Wat trouwens ook raar was, ik weet niet waarom, maar volgens mij heb ik de verkeerde producten onthouden.
O Hoe bedoel je?
R In mijn beleving stond er eerst wat anders, maar zal t wel fout hebben.
O Owh, apart. Ik zal er een notitie van maken.

19. Respondent (existing – swap) – Male – Spoke Up

R Er was wat raars aan de hand.
O Wat was er aan de hand?
R Ik heb producten onthouden en toen ik terugkwam stonden er twee andere op. Heb jij ze tussendoor gewisseld of zoiets?
O Ja, dat klopt.
R Dat dacht ik al, was dat ook het doel van het onderzoek?
O Klopt (en dan leg ik het hele doel van het onderzoek uit…). Kan jij je redenen bedenken waarom andere respondenten dit niet tegen mij zouden zeggen?
R Om er makkelijk vanaf te zijn, jij kan namelijk niet controleren wat zij wel of niet gezien hebben. Het zou kunnen dat ze de swap niet zien, maar dan heb je de lijst ook niet serieus ingevuld. Het viel mij behoorlijk op. Maar goed als je eerstejaars bent kan het ‘te eng’ zijn om een onderzoeksleider aan te spreken.
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Je had nog meer kunnen benadrukken dat je ethisch verantwoord gedrag verwacht van de respondenten. Dan kunnen ze niet ontkennen dat ze verkeerd handelen.

Disconfirmed Expectancies & Exposure to Misinformation Condition (*n* = 5 | 3 interviews + 2 spoke up)

4. Respondent (non-existing – swap) – Male

O En hoe vond je dat het ging?
R Ik hoop dat ik alles goed gedaan heb! Het was wel moeilijk.
O Hoe dat zo?
R Ik had die producten allemaal onthouden dacht ik. Kom ik in de winkel, kan ik er twee niet vinden. Kom ik hier terug, blijkt dat ik de verkeerde heb onthouden. Ik begon wel erg aan mezelf te twijfelen, wat nu goed was en niet. Moet ik anders nog een keer terug om die andere producten nog te doen?
O Nee dat hoeft niet, goed dat je het aangeeft in ieder geval!
R Ja sorry, ik ben normaal wel goed in dingen onthouden weet ook niet wat er fout ging. Hoop dat je nog wat aan de data hebt.
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Nee

8. Respondent (non-existing – swap) - Female

O En hoe vond je dat het ging?
R Slecht, ik heb maar wat gedaan.
O Hoe komt dat?
R Vond het vet moeilijk om alles te onthouden, was de helft alweer vergeten toen ik in de winkel stond.
O Waar komt dat door denk je?
R Vind het sowieso moeilijk om zoveel dingen te onthouden, maar als ik helemaal eerlijk ben had ik ook wel iets beter kunnen leren.
O Hoe bedoel je?
R Ik heb die producten een paar keer doorgelezen en toen dacht ik dit gaat wel lukken. Had het een beetje onderschat denk ik. En toen kwam ik terug bleek ik ook nog de verkeerde Taksi onthouden te hebben. Toen dacht ik, laat maar. Ben benieuwd hoeveel ik er goed heb.
O Om objectief te blijven weet ik de exacte prijzen ook niet.
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Wat gaat er met deze data gebeuren? Komt er te staan dat ik het heel slecht gedaan heb?
O Alles wordt anoniem behandelt, niemand komt er achter wie er precies wat goed/fout gedaan heeft.

12. Respondent (non-existing – swap) – Spoke Up - Female

R Dit was lastig zeg! Ik dacht dat doe ik even, maar toen ik in de winkel stond moest ik echt even graven wat de producten ook weer precies waren. Was dat ook een onderdeel van het onderzoek?
O Het was ook gedeeltelijk om je geheugen te testen ja.
R Zoiets dacht ik al ja, het was moeilijk.
O Je hebt er qua behoorlijk lang over gedaan, zowel het bezoek als het invullen van de vragenlijst. Hoe komt dat?
R Ik had in de winkel moeite alle producten te vinden.
O Maar heb je uiteindelijk alles wel kunnen vinden?
R Ja.
O En het invullen?
R Ja ik zal het wel verkeerd onthouden hebben, maar had twee andere producten. Heb maar gewoon de prijs opgeschreven die ik dacht dat het zou hebben.
O Ook als je weet dat dit wellicht niet de goede prijs is?
R Ja ik dacht ik zet maar wat neer, anders lijkt het ook zo alsof ik helemaal niks kan onthouden.
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Ik zou minder producten doen, zes was echt teveel voor mij.

15. Respondent (non-existing – swap) – Spoke Up - Male

R Je hebt het me wel moeilijk gemaakt hoor. Mijn hoofd duizelt nog een beetje na, de prijs voor koffie ga ik niet meer vergeten.
O Maar het is allemaal gelukt?
R Ja volgens mij wel, heb alleen wat producten door elkaar gehaald denk ik.
O Hoe dat zo?
R In mijn beleving stond er eerst wat anders.
O Hoe bedoel je?
R Ik had twee andere producten in m’n hoofd toen ik winkel in ging, maar bleek wat anders te zijn. Zal het wel verkeerd onthouden hebben. Hoop dat dit nog wel bruikbaar is?
O Goed dat je het zegt, ik zal binnenkort nog eens kritisch naar de lijst kijken.
R Ja, lijkt me handig.
O Zou jij redenen kunnen bedenken om dit soort dingen niet tegen mij te zeggen? Niet iedereen komt er namelijk zo eerlijk voor uit.
R Om er makkelijk vanaf te zijn. Ik twijfelde al om je eerder te roepen, maar toen dacht ik dat is alleen maar gedoe om iets wat ik fout heb gedaan. Daar wilde ik je niet mee storen. Het is dat nu dit interview is anders betwijfel ik of ik het gezegd zou hebben.
O Kun je dat uitleggen?
R Als ik dit niet gezegd zou hebben dan had jij niet geweten dat ik de verkeerde producten had onthouden.
O Dat is waar, goed dat je zegt. Iets voor in de discussie.
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?

R Ik zou zoveel mogelijk respondenten, het liefst iedereen, om te vragen hoe het ging. Anders kan er wellicht nuttig informatie achter wegen blijven.
O Maar is dat ook niet de verantwoordelijkheid van de respondent?
R Zo zou je het ook kunnen zien, maar het speelt wel vals spelen in de kaart.

18. Respondent (non-existing – swap) - Female

O En hoe vond je dat het ging?
R Ging wel goed denk ik, was wel even flink nadenken zeg.
O Denk je dat je alles goed gedaan hebt?
R Dat ligt er aan wat het doel van het onderzoek was. Ging het echt alleen om die producten?
O Hoezo, denk je dat er meer aan de hand is?
R Weet ik niet, vond het in ieder geval lastig!
O Heb je doorgehad dat je vragenlijst tussendoor gewisseld is?
R Dus toch, ik wist het! Ik dacht al dat ik gek geworden was. Ik heb er nog naar gezocht maar kon niet vinden hoe ‘t gedaan moest zijn. Maar dan heb ik denk ik al nu al fout gehandeld. Ik heb de prijzen opgeschreven van de andere producten. Krijg ik nu een onvoldoende? Klopt het trouwens dat die eerste producten niet te vinden waren of heb ik slecht gezocht?
O Nee dat klopt.
R Gelukkig, anders kon ik er écht helemaal niks van.
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Ik vertrouw geen enkel onderzoek meer vanaf nu.

Control Condition (*n* = 4 | 3 interviews + 1 spoke up)

1. Respondent (existing – no swap) – Male – Spoke up

R Het was lastig om die zes producten allemaal te onthouden, maar als je even de tijd neemt lukt het wel. Ik had er een verhaaltje van gemaakt. Sommige waren nog wel moeilijk te vinden in de winkel. Het tweede gedeelte was duidelijk moeilijker. Eerste gedeelte was gewoon rustig een hamburger halen, hoefde je niet echt bij na te denken.
O Welke producten waren lastig te vinden voor je?
R Het duurde even voordat ik doorhad dat de chocomel niet bij de frisdrank in de buurt stond. Ook was de campina lastig te vinden. Maar uiteindelijk is alles gelukt toch?
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Nee, was leuk onderzoek.

5. Respondent (existing – no swap) - Female

O En hoe vond je dat het ging?
R Ging redelijk goed volgens mij. De eerste keer was wel even spannend omdat je denkt dat iedereen je door heeft, maar de tweede had ik daar al geen last meer van. Ik ben trouwens de bon vergeten van de chips net, is dat erg?
O Nee. Is het gelukt alle prijzen te onthouden?
R Dat was wel even diep nadenken net, maar ik heb ze er ingestampt in de winkel. Gewoon het rijtje opdreunen. Heb ze volgens mij wel allemaal goed.
O Andere respondenten gaven aan dat zij het wat moeilijker vonden, kan je dat begrijpen?
R Als je snel de producten doorleest en dan vlug de winkel in gaat om er maar vanaf te zijn dan snap ik dat het moeilijk is. Maar als je even de tijd er voor neemt dan is het goed te doen. Er staat toch meer dan genoeg tijd voor.
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Valt het niet erg op met die tas? Ik bedoel, er komen zoveel mensen met dezelfde tas binnen dat moet opvallen toch?

9. Respondent (existing – no swap) - Male

O En hoe vond je dat het ging?
R Goed! Die hamburger zag er trouwens wel een beetje raar uit, het bakje plakte ook helemaal. Vond dat niet zo netjes van ze. Gebruiken ze altijd dat bakje? Ik kom hier vaker, maar heb dit nog niet eerder gezien.
O Dit is standaard bij de verkoop van vleeswaren bij de slagerij. En hoe ging het tweede gedeelte van het onderzoek?

R Dat was wel wat lastiger, al die prijzen en producten lijken op een gegeven moment op elkaar. Ben echt wel drie keer teruggelopen naar de koffie.
O Maar was het wel te doen?
R Ja op zich wel, het lijkt moeilijker dan dat het is. Ik weet niet of ik alle prijzen goed heb, twijfelde nog over de Peijenburg. Ik denk dat ik deze prijzen nooit meer ga vergeten (noemt t rijtje nogmaals helemaal op).
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Was echt leuk om een keer te doen, een keer heel wat anders. En het viel me op dat de manager naar me keek met zo'n blik van herkenning, dat was wel grappig.

13. Respondent (existing – no swap) - Male

O En hoe vond je dat het ging?
R Geen probleem, dit was goed te doen. Ik heb eerder wel eens mee gedaan en dan moest je echt drie A4 onthouden, dus dit was prima. Ik dacht in het begin wel dat t moeilijker zou zijn met al die prijzen, maar als je dat dan even in je hoofd stampt en je loopt twee rondjes door de supermarkt is het zo gedaan.
O Andere respondenten gaven aan dat zij het wat moeilijker vonden, kan je dat begrijpen?
R De ene persoon zal altijd dingen beter kunnen onthouden dan andere. Als je daar moeite mee hebt kan ik me indenken dat het lastig voor je is. Maar aan de andere kant, ik ben zelf ook geen ster in onthouden (note: hij was de eerste afspraak vergeten) en het is mij ook gelukt. Denk dat het voornamelijk te maken hebt met hoe serieus je dit neemt.
O Hoe bedoel je?
R Als je dit even tussendoor doet, alleen voor de punten bijvoorbeeld, dan kan je gewoon wat invullen en weer weggaan. Er is weinig controle snap je.
O Zijn er nog dingen die je kwijt wilt, dingen die je opvielen, verbeterpunten?
R Hele lelijke tas

APPENDIX II: Checklist

Thirty-two items divided over five categories**:**

Characteristics of the mystery shopper [3]:

- age [-open-]
- gender [M/F]
- experience with mystery shopping [none/little/somewhat/much/greatdeal]

Characteristics of the visit [7]:

- day of the week [M/T/W/T/F/S/S],
- timeframe for the visit [8-10am/10-12am/12am-2pm/2pm-4pm/4pm-6pm/6pm-8pm/8pm-10pm],
- n of customers [0-10/10-25/25-40/>40],
- n of cash desks total [1-6],
- n of cash desks open [1-6],
- n of customers in line in front of mystery shopper [-open-],
- n of customers in line behind mystery shopper [-open-]

Interior [15]:

- price of filler product A (coffee, Douwe Egberts Aroma Rood Snelfiltermaling 500 gram) [-open-],
- price tag filler product A clean [Y/N],
- damaged products in shelf of filler product A [Y/N],
- price of filler product B (sandwich filling, De Ruijter Chocoladehagel Puur 400 gram) [-open-],
- price tag filler product B clean [Y/N],
- damaged products in shelf of product filler B [Y/N],
- price of existing target product (soft drink, Taksi Tropisch Fruit 1.5 liter) [-open-],
- price tag existing target product clean [Y/N],
- damaged products in shelf of existing target product [Y/N],
- price of non-existing target product (soft drink, Taksi Bosvruchten 1.5 liter) [-open-],
- price tag non-existing target product clean [Y/N],
- damaged products in shelf of non-existing target product [Y/N],
- price of filler product D (dairy product, Campina Vlavlip Vanille 1 liter) [-open-],
- price tag filler product D clean [Y/N],
- damaged products in shelf of filler product D [Y/N],
- price of filler product E (chocolate milk, Chocomel Vol 1 liter) [-open-],
- price tag filler product E clean [Y/N],
- damaged products in shelf of filler product E [Y/N].

Familiarity with products [5]:

- familiarity with filler product A [strongly agree/agree/undecided/disagree/strongly disagree]
- familiarity with filler product B [strongly agree/agree/undecided/disagree/strongly disagree],
- familiarity with target product [strongly agree/agree/undecided/disagree/strongly disagree],
- familiarity with filler product C [strongly agree/agree/undecided/disagree/strongly disagree],
- familiarity with filler product D [strongly agree/agree/undecided/disagree/strongly disagree],

Evaluation [2]:

- overall grade of supermarket [1-10],
- explanation [-open-].

APPENDIX III: Instruction Document

The instruction document contained an act of confidentiality and a procedure:

Act of Confidentiality

De data zal **anoniem** verwerkt worden. In verband met privacy- en copyrightrechten mogen er **geen beeld- en/of geluidsopnamen** (foto’s/video’s) van dit document, de winkel en de producten gemaakt worden.

Door onderstaande gegevens in te vullen geef ik aan dat ik vrijwillig deelneem aan dit onderzoek, maar behoud ik het recht om te allen tijde te kunnen stoppen met dit onderzoek:

Naam	
E-mail	
Handtekening	

Protocol

Briefing

In dit onderdeel zult u een bezoek, als undercover klant, gaan brengen aan de **gehele Emté supermarkt**. Het is bij dit onderdeel belangrijk om te onthouden dat in tegenstelling tot Deel I er **absoluut geen interactie met medewerkers** gezocht mag worden! De Emté wil een klantbeleving meten die zich richt op de winkel en niet op het personeel. Ook is het gebruik van smartphones in en rondom de winkel verboden. Voorbeelden van onderwerpen waar op gelet wordt zijn **de netheid van de winkel**, de **aanwezigheid en prijs van producten** en de **vulgraad van de schappen**. Lees de checklist (volgende pagina) goed door en onthoud waar u op moet letten. Gaat u verder met: H*et Bezoek*.

Vergeet u niet een handtekening te zetten op het voorblad voordat u het bezoek doet. Hiermee geeft u aan dat u het doel van het onderzoek snapt en dat u vrijwillig mee doet.

Het Bezoek

Denk er aan, mystery shoppers zijn en blijven altijd anoniem! Gedraagt u zich daarom zoveel mogelijk als een gewone klant. U draagt te allen tijde de tas, hieraan kan de manager zien dat u een mystery shopper bent.

De route door de winkel staat vrij, u loopt er doorheen zoals u wilt en neem zoveel tijd als u nodig heeft. Als u alle items van de checklist geobserveerd heeft kunt u de winkel verlaten. Voor de Emté zijn vragen over **de producten** vooral belangrijke aandachtspunten. Koop *een zak chips* en verlaat de winkel. Vergeet niet het bonnetje mee te vragen. Ga door naar: *De Debriefing*.

De Debriefing

Na het verlaten van de winkel gaat u weer terug naar het huis van de onderzoeker (Kuipersdijk 132) om de vragenlijst in te vullen en een korte debriefing te krijgen van het onderzoek. Een random geselecteerde groep deelnemers aan het onderzoek dient nog deel te nemen aan een kort follow-up interview (5 – 10 minuten) over de bevindingen. De onderzoeker zal tijdens de debriefing laten weten of dit u betreft.