Strategy and Misdirection in Forced Choice Memory Performance Testing in Deception
Detection

Robin Orthey

Department of Psychology

University of Portsmouth

Portsmouth, UK

Aldert Vrij

Department of Psychology

University of Portsmouth

Portsmouth, UK

Sharon Leal

Department of Psychology

University of Portsmouth

Portsmouth, UK

Hartmut Blank

Department of Psychology

University of Portsmouth

Portsmouth, UK

Abstract

We examined forced choice memory performance testing in deception detection from a theoretical perspective. Evidence suggests that participants form different strategies to defeat this test. We attempted to describe these strategies within the framework of Cognitive Hierarchy Theory (CHT), a theory that distinguishes strategies based on their degree of anticipation of opponents' strategies. Additionally, we explored whether the strategy selection process is malleable. Truth tellers and liars were subjected to a forced choice memory test about a mock crime. Additionally, half of the sample was subjected to a misdirection changing the appearance of the test to that of a polygraph examination. We found detection accuracies and strategies similar to previous experiments and our misdirection manipulation elicited new strategies and behaviour. Theoretical and practical applications are discussed.

The field of deception detection currently features two general approaches. One is the elicitation of verbal, non-verbal, or physiological cues to deception, that allow to differentiate truth from lies. The other is the detection of concealed information. The idea of the latter approach is to detect knowledge in suspects that only the investigators and the perpetrator are aware of. Involvement is concluded by inference. If the suspect has intimate crime knowledge then s/he must be somehow involved in the crime. In this article we focus on the Symptom Validity Testing (SVT).

SVT started as a malingering detection tool for fake cognitive impairment (Pankratz, Fausti, & Peed, 1975). Pankratz et al. (1975) describe a case of alleged loss of hearing. They presented their examinee with a sound in one of two temporal intervals and asked which of the two contained the sound. The examinee was instructed to indicate the correct interval or guess if he did not know. This process was repeated over 100 of similar trials. In their case, the client indicated 36 out of 100 times the correct time interval. The probability of having only 36 answers correct is smaller than .004. Therefore the authors considered the loss of hearing to be malingered. The idea behind this method is that genuine performance, that is impaired hearing capabilities, would force the examinee to guess on each trial. Consequently, the total test score is expected to fall within levels of chance. The authors infer malingering from underperformance, that is test performance worse than expected by chance, as a sign of deliberate avoidance of correct answers.

Since then, a variety of SVT tools have been developed, but the core principle as described in Pankratz et al., (1975) remains the same throughout. In cases of deception detection or fake memory loss an event specific binary forced choice memory performance test is used (Bianchini, Mathias, & Greve, 2001; Van Oorsouw, & Merckelbach, 2010). Examinees are presented with questions about the event and a pair of answer alternatives. One alternative is always correct, the other alternative is always incorrect. Liars are expected

to display underperformance (because they recognize the correct answer and purposefully select the incorrect answer), while truth tellers, who have no knowledge of the event, are expected to score within levels of chance (because they actually guess). Empirical studies report a high (90-100%) classification rate for truth tellers (specificity) (Giger, Merten, Merckelbach, & Oswald, 2010; Meijer, Smulders, Johnston, & Merckelbach, 2007; Shaw, Vrij, Mann, Leal, & Hillman, 2012), and a moderate detection rate (40-63%) for liars (sensitivity) (Giger et al., 2010; Jellicic, Merckelbach, & van Bergen, 2004; Meijer et al., 2007; Merckelbach, Hauer, & Rassin, 2002; Shaw et al., 2012). In other words, 90 – 100% of truth tellers typically perform at chance levels, whereas 40 – 63% of liars typically underperform. Overperformance – total scores better than chance – are currently not interpreted as diagnostic in forced choice memory deception detection.

A major problem of the field is that little attention has been paid to the theoretical background of forced choice memory performance testing. Liars' avoidance behaviour has been assumed but not explained. Exceptions are Shaw et al. (2012), who refer to a general avoidance tendency found in interviewing literature; and Meijer et al. (2007) who, apart from this avoidance tendency, also argue that examinees may fail the test due to their inability to produce genuine randomness. It seems that the generally accepted underlying mechanism of the forced choice performance tests is an avoidance preference of true crime information by liars. This theoretical concept can explain why the test detects liars, but it cannot explain why a considerable proportion of liars (often more than 50%) are *not* detected. Here we propose and explore a new theoretical perspective on forced choice memory performance testing, which is also capable of predicting cases that avoid detection.

In their study, Shaw et al. (2012) also presented the self-reported strategies of their participants. For liars, these included countermeasures to appear innocent, such as 'avoiding correct information', 'deliberately choosing incorrect answers', or 'motivated randomisation'.

The latter strategies suggest an understanding of the test's mechanism, as the authors noted themselves. In addition, they found that participants who understood the test's rationale were also more likely to avoid being detected. Similarly, Verscheure, Meijer, and Crombez (2008) obtained the same effect when they compared coached liars (who were informed about the working of the test) with naïve liars. Coached liars escaped detection, while naïve liars were detected with the same accuracy as reported in other studies. This suggests that liars' test behaviour is a product of their strategy and understanding of the test's mechanism, which would not only explain why some liars are detected, but also why some are not detected by the test.

One theory suited for analysing strategies in forced choice performance testing is Cognitive Hierarchy Theory (CHT; Camerer, Ho, & Chung, 2004). According to this theory a strategy can vary in its level of sophistication, which is the degree it accounts for an opponent's strategy. These degrees are expressed in numerical levels. In this case, a level 0 strategy does not consider how the test tries to identify the examinee and the examinee may just comply with the test's instructions ('Select the correct answer, if you don't know it guess.'). A level 1 strategy would be based on the idea that the test identifies the guilty through their compliance to test instructions and therefore choose countermeasures that work against these instructions (such as e.g. 'deliberately avoiding correct information'). Subsequently, a level 2 strategy would assume that the test expects a level 1 strategy and therefore it consists of countermeasures to counter a level 1 strategy, for example to 'deliberately include correct information' or 'making responses look random'. Theoretically, there is no limit to the strategy level, but a key feature of CHT is that the process of strategy selection is limited by the cognitive resources of the examinee. Camerer et al. (2004) refer to an average level of 1.5, which means that the majority of people will either form a level 1 or

2 strategy. Thus, we conceptualize suspects' behaviour in forced choice memory tests in terms of the sophistication of their chosen strategy.

Given the assumption that understanding and strategy selection are crucial to the test's detection efficiency (Shaw et al., 2012; Verscheure et al., 2008), we explore two questions. First, we will examine the role of strategy selection, as defined in CHT, in relation to detection efficiency. To do so, we will measure the examinee's self-reported strategies, translate them into CHT terms, and examine which strategies the test detects and which not. We formulate the following hypotheses: (1) Liars who use level 1 strategies will be detected, but liars who use level 2 strategies will not be detected; (2) Liars will report higher order strategies than truth tellers; and (3) Specifically, we expect liars average strategy level to be higher than zero, but not truth tellers', because they are assumed to comply with the test's instructions and guess.

Second, for two reasons we will investigate whether it is possible to influence the strategy selection itself. On the one hand liars not only need to behave differently from truth tellers, but their behaviour as a group must also be homogenous to ensure reliable detection accuracy. Shaw et al. (2012) demonstrate that liars choose from a multitude of strategies, but the test is only designed to detect one of them (avoiding correct information). On the other hand, if we can influence the strategy selection process we can attempt to elicit new behaviours in liars that subsequently can be used for detection purposes. One example is overperformance, which is currently not conceptualized in deception detection, but it shares the same properties as underperformance. Truth tellers will exhibit overperformance through chance, but liars are just as able to produce over- as underperformance (each requires the liars to recognize the correct answer). To elicit overperformance in liars we will utilize a misdirection of reasoning (Kuhn, Caffaratti, Teszka, & Rensink, 2014). By attaching half of our sample to a fake skin conductance response (SCR) sensor we intend to create the

impression of a polygraph examination. This manipulation is based on the widespread believe that deception can be inferred from physiological signals. Since the SCR sensor is a very salient part of the test situation we expect it to act as a mask for the actual mechanism of forced choice memory performance testing. If examinees mistakenly believe that classification takes place through physiological measures, they are more likely to comply with the test's instructions and actually select the correct answers, or only perform countermeasures against the physiological measurements.

Here we attempt to elicit overperformance and formulate three hypotheses: (4) We expect our misdirection manipulation to decrease the likelihood that liars and truth tellers realize the actual classification mechanism of forced choice memory performance testing; (5) We expect examinees in the misdirection condition to use physiological countermeasures as their strategy to beat the test; and (6) We expect liars in the misdirection condition to produce more cases of overperformance (significantly more questions correct than expected by chance) than liars in the control condition.

Method

Participants

A total of 95 undergraduate students and members of staff of the University of Portsmouth participated in this study. Three participants were excluded from the analysis, because they were familiar with the mechanism of forced choice testing or did not follow the instructions. The final sample consisted of 92 participants (37 male & 55 female, *mean* age = 25.45, *SD* = 9.66). The experiment was approved by the ethics committee of the University of Portsmouth.

Material

An assumption of the forced choice memory performance testing is that the answer alternatives are equally plausible so that truth tellers (those who do not know the correct answers) will consider both answer alternatives for each question equally likely to be correct (Bianchini et al., 2001; Doob, & Kirschenbaum, 1973). We constructed 23 questions pertaining to the mock crime procedure. These 23 questions, with two answer alternatives each, were then subjected to a pilot procedure to ensure that the answer alternatives were equally plausible. In this pilot participants were blind to the mock crime and presented with the questions and answer alternatives. They were tasked to indicate for each question the answer they thought was the most plausible. A set of answer alternatives was deemed plausible when one option was not more frequently chosen than 70% (just as in Jelicic et al., 2004; Merckelbach et al., 2002). In total, four pilot cycles (N = 24/20/20/21) were required to find for each question an equally plausible pair of answers.

In total, twenty questions featured verbal answer alternatives and three questions featured pictorial answer alternatives. Pictures were taken from the Psychological Image Collection at Sterling (PICS; Hancock, 2014) face database.

Procedure

Participants were informed that they had to beat a lie detection test over a warehouse burglary. They were rewarded with either course credit (first year undergraduate students) or a £5 voucher (other participants). Additionally, they had the opportunity to win one of two £50 vouchers if they were able to appear innocent on the lie detection test.

Participants were randomly assigned to either a mock crime or an innocent condition. In the mock crime condition the participant (liar in the subsequent test) had to plan and execute a mock burglary. This burglary scenario was completed on a computer. To make the burglary task more meaningful and memorable for participants, textboxes were provided that

described the different situations and the participant was asked to make key decisions throughout the scenario (e.g. ‘What kind of product would you like to steal?’ Answer: A: Laptop B: Tablet). Furthermore, each option was presented with an advantage and disadvantage that was related to an increase or decrease of profit and safety (e.g. for option A: Laptop Advantage: *very valuable*, Disadvantage: *big*). The chosen options were subsequently considered as the ‘truthful’ options during the test procedure later on (and thus could differ for each participant). Next, a 5 minutes filler task (short personality questionnaire) was implemented in order to have a break between mock-crime and lie detection test, because we were concerned that the test’s rationale would be too obvious if the test was conducted directly after the mock crime.

In the innocent condition, participants (truth tellers in the subsequent test) did not perform the mock crime, but just the filler task.

Participants (liars and truth tellers) were then informed that they were suspected of a burglary in a police investigation and that they would be submitted to a lie detection test. Half of the participants were attached to a fake SCR sensor and led to believe that their sweat production during the test would be monitored (the other half was not attached to anything nor any information was given). This factor is labelled ‘Misdirection’. Participants were told that during the lie detection test they would be presented with questions about the burglary and two answer alternatives. It was their task to indicate the correct answer and, in case they did not know it, guess.

A total of 23 questions were presented in two steps. First, a question was presented. Once read, the participant could move on to a new window, where both answer alternatives were presented next to each other horizontally. The horizontal alignment was determined

randomly. The order in which questions were presented was counterbalanced using a latin square of the size 23.

After the test participants were notified that the lie detection test was over and were asked to answer the following questions honestly: ‘What did you do to appear innocent on the test?’ and ‘Did you believe that your sweat was measured during the test? (Yes/No)’. These questions were used to determine the strategy each participant used and to check whether participants in the Misdirection condition were misdirected by the fake SCR sensor.

Finally, liars were again shown the 23 test questions. Liars were instructed to indicate the correct answer, which served as a memory check.

Design

This study used a double-blind design and participants were assigned to a condition by the computer. It featured a 2 (Veracity: lie vs. truth) x 2 (Misdirection: yes vs. no) between subjects design with the deviation from chance performance as dependent measure. Deviation from chance performance was expressed unidirectionally (only underperformance as criterion) and bidirectionally (under- and overperformance as criterion). First, we computed the z -score for each participant’s total test score using Siegel’s (1956) formula for binomial distributions. Negative scores indicated tendencies towards underperformance and positive score towards overperformance. These scores were used for unidirectional testing. For bidirectional testing we used the absolute version of these scores. In this case the larger a score the more did the responses show either under- or overperformance. Z -scores were chosen over raw test scores, because they are independent from the total number of questions asked and by definition indicate how much the score deviates from the chance distribution.

Detection accuracy is expressed in terms of sensitivity (, the likelihood that a guilty participant is correctly detected), specificity (, the likelihood that an innocent participant is

correctly detected,) and the Area Under the Curve (AUC), which is the general detection accuracy for the entire scale. Sensitivity and specificity require a specific cut off. However, the choice of cut off is under debate (e.g. Binder, Larrabee, & Millis, 2014). For comparison with other deception detection experiments we report sensitivity and specificity utilizing the commonly used 5% cut off. Scores equal to or smaller than -1.65 unidirectionally and scores larger or equal to 1.65 bidirectionally were considered deceptive. 95% confidence intervals were provided with square brackets.

Participants' strategies were extracted from the open question 'What did you do to appear innocent on the test?' The primary investigator first read through all responses and then classified them into the following eight categories: (1) *No strategy* represents examinees who reported answering the questions honestly or reported having no strategy. (2) *Avoiding correct information* refers to responses that indicate that all correct answer alternatives were deliberately avoided. (3) *Mixture of truth & lies* indicates that the participant deliberately included correct and incorrect answer in his/her response scheme. (4) *Imitating ignorance* refers to cases where the participant either states to imitate a response pattern of a truth teller or make his/her answering pattern look random. (5) *Deductive guessing* represents answers that indicate selecting the most obvious or logical answer. (6) *Demeanor* refers to cases where respondents control their facial expressions or body posture. Finally, (7) *Physiological countermeasures* represents strategies directed at disrupting physiological measurements, such as breath control or moving ligaments that were attached to the fake SCR sensor. Answers that did not address the question or made no sense were indicated as (8) *Other* and excluded from further analysis.

A second rater, blind to the hypothesis, classified each participant according to these eight categories. If a response would have fitted into more than one category, the coder was instructed to choose the one with the best fit. In cases of disagreement both coders discussed

the instance and coded the case independent from each other again. Inter-rater reliability was good (73.9% absolute agreement).

Subsequently, we created a new variable that indicated each strategy level according to CHT criteria. We defined three levels (0 – 2). Level 0 strategies (1) represent simple compliance with the test instructions. Level 1 strategies (2,6,7) represent participants' reaction to the test instructions or situation (e.g. 'Avoiding correct information' or 'Controlling non-verbal behaviour'). Level 2 strategies (3,4,5) were defined as reactions to level 1 strategies (e.g. 'random responding'). Inter rate reliability was good (83.7% absolute agreement).

Two variables were created that described the participants' beliefs about the method underlying the test. The first was a binary indication of whether or not the participant understood that too many incorrect answers would identify them as liars. Both the primary investigator and a blind rater used the question 'What did you do to appear innocent on the test?' to make this judgement. In cases of disagreement both coders discussed the instance and coded the case independent from each other again. Inter-rater reliability was very high (97.8% absolute agreement). The second variable indicated whether the participant believed that their physiological responses were measured. Participants indicated their response on the question 'Did you believe that your sweat was measured during the test? (Yes/No)' during the procedure.

Lastly, we computed a measure for the memory of event information. The memory rating was produced for liars and was the sum of correct answers indicated during the memory check at the end. Memory of correct crime information was high (*mean* = 81.66, *SD* = 10.6).

Results

Understanding & Misdirection

First, we examined the effects of our Misdirection manipulation. We expected our misdirection condition to decrease the likelihood of understanding the true mechanism of the SVT (H4).

First, we checked whether participants in the misdirection condition did actually believe that their physiological responses were measured. Of the 23 liars allocated to this condition, 82.6% believed the misdirection, while 95.65% of the 23 truth tellers allocated to this condition did so, which suggests that our manipulation was convincing.

We then checked whether our misdirection manipulation affected the likelihood of a participant to understand the underlying rationale of the lie detection test. We found no difference in liars' ability to discern the test's mechanism, $X^2(1, N = 46) = 1.075, p = .299$, between Control (35%) and Misdirection (22%) condition. For truth tellers the misdirection manipulation greatly reduced the likelihood to discover the test's mechanism: 30.4% of the truth tellers in the control condition understood how the test works, whereas nobody in the misdirection condition did, $X^2(1, N = 46) = 8.256, p = .014$. This supports Hypothesis 4 only partly, as we expected both liars and truth tellers to display a decreased likelihood of discerning the test's classification mechanism.

Strategies

Next, we will give an account of the strategies that participants used and explore differences in strategy levels. We expected our Misdirection condition to elicit reports of physiological countermeasures (H5). In terms of strategy levels we expected that liars used more sophisticated strategies than truth tellers (H2) and that liars' strategies were more sophisticated than level 0 strategies (H3). Then we will address the detection accuracy of the

different strategy levels. We expect the test to reliably detect level 1 strategies, but not level 2 strategies (H1).

Table 1 lists the frequencies of strategies broken down by Veracity and Misdirection. For truth tellers the most prevalent strategy was to have either no strategy or just to be honest (30.4 and 39.1% in the two Misdirection conditions – Control and Misdirection – respectively). Some truth tellers indicated to deliberately imitate ignorance (13%) or to pick the most logical answers (Deductive guessing: 21.7 and 13%). For liars several popular strategies emerged. Avoiding correct information (20 and 30.4%), providing a mixture of correct and incorrect answers (21.7%) and imitating ignorance (30.4 and 17.4%) were the most popular strategies.

The Misdirection and control Conditions differed most from each other in a strategy that is unique to each condition (for liars and truth tellers alike). In the Control condition 17.4% of truth tellers and 8.7% of liars reported controlling their demeanor during the test. In contrast in the Misdirection condition around 21.7% of the truth tellers and 17.4% of the liars reported countermeasures that were directed against our fake SCR sensor (physiological countermeasures). The presence of self-reported countermeasures against physiological sensors in the Misdirection condition supports H5.

TABLE 1 HERE

Next, we examined the strategy levels. A 2 (Veracity) x 2 (Misdirection) between-subjects ANOVA was conducted with the strategy level as dependent variable. There was a significant difference for the Veracity main effect, $F(1,82) = 10.65$, $p = .002$, $\eta^2 = .12$. Liars ($mean = 1.50$ [1.28 1.72]) used on average a higher level strategy than truth tellers ($mean = 0.98$ [0.75 1.20]), supporting H2. The main effect for Misdirection, $F(1,82) = 1.02$, $p = .315$,

$\eta^2 = .01$, and the interaction between Veracity and Misdirection, $F(1,82) = 0.02$, $p = .904$, $\eta^2 < .001$, were not significant.

Subsequently, we tested whether the strategy level would differ from zero. One-sample t-tests with the strategy level as dependent measure were conducted for the liars and truth tellers in the Control and Misdirection conditions. Liars' average strategy level significantly exceeded zero in the Control condition, $t(21) = 12.64$, $p < .001$, $mean = 1.59$ [1.33 1.85], and the Misdirection condition, $t(21) = 11.20$, $p < .001$, $mean = 1.41$ [1.15 1.67]. Unexpectedly, truth tellers' average strategy level did exceed zero in the Control condition, $t(20) = 5.55$, $p < .001$, $mean = 1.05$ [0.65 1.44], and Misdirection condition, $t(20) = 4.66$, $p < .001$, $mean = 0.90$ [0.50 1.31]. Therefore, H3 is only partly supported.

In Table 2 the percentage of detected and undetected liars is displayed for the Control and Misdirection condition. Due to the fact that only two observations were available for level 0 strategies we forfeited any interpretation. For level 1 strategies we found a high detection rate in our Control (around 85%) and Misdirection condition (around 72%). For level 2 strategies we found the same results in both conditions, half of the liars who used level 2 strategies were detected. In line with Hypothesis 1 we found a high detection rate of level 1 strategies in liars. Contrary to our expectations half of the liars with level 2 strategies were also detected. Thus, H1 is only partly supported.

TABLE 2 HERE

To follow up we examined in particular which level 2 strategies of liars exactly were detected and which remained undetected. Table 3 displays these frequencies for the Control and Misdirection condition together, as both showed almost the same pattern. As Table 3 shows, each of three level 2 strategies was as frequently detected as it remained undetected. In addition we looked into the individual z-scores of these participants. A considerable

proportion (33.33%) of detected liars using level two strategies had just enough answers wrong to be detected.

TABLE 3 HERE

Avoidance behaviour & Detection accuracy

Lastly, we examined the detection accuracy. We expected to find greater overperformance in the Misdirection condition (H6; overperformance is incorporated in the bidirectional criterion).

In Table 4 we summarize the detection parameters for the Control and Misdirection condition for both uni- and bidirectional avoidance behaviour. Sensitivity and specificity are high in every case. The traditional approach (Control – unidirectionally) obtained a sensitivity of 56.52% and a specificity of 86.95%. With a bidirectional decision criterion we achieved even higher sensitivity (65.22%) and specificity (95.65%) in the Misdirection condition utilizing the bidirectional criterion. In terms of generalized detection efficiency uni- ($AUC = .76, p = .002$) and bidirectional ($AUC = .72, p = .011$) classification was significantly better than chance in the Control condition. In the Misdirection condition, only the bidirectional ($AUC = .82, p < .001$), but not the unidirectional ($AUC = .67, p = .055$), measure provided better discriminative ability than chance. This supports H6, as only the bidirectional criterion (includes overperformance) and not the unidirectional criterion was significantly better than chance performance.

Finally, we examined what strategies were used by liars in the Misdirection condition exhibiting overperformance. Of the five cases of overperformance, the categories ‘No Strategy’ and ‘Mixture of truth & lie’ were reported by one participant and three reported performing ‘Physiological countermeasures’.

TABLE 4 HERE

Discussion

The aims of this study were twofold. First we attempted to theoretically conceptualize forced choice memory performance testing in terms of strategy selection processes. We defined strategy selection in terms of Cognitive Hierarchy Theory (Carmerer et al., 2004). Key concepts of CHT involve differentiation between levels of strategies through the degree of anticipation for opponents' strategies and the limitations imposed by individual cognitive capacities in the strategy selection process. Second, we investigated the malleability of the strategy selection process: Through a misdirection of reason (Kuhn et al., 2014) we attempted to elicit cases of overperformance in liars.

Overall we found a relatively high detection accuracy for liars (56.2 – 65.22%) and truth tellers (86.95 – 95.65%). This matches sensitivity and specificity generally found in previous experiments (Giger et al., 2010; Jelcic et al., 2004; Meijer et al., 2007; Merckelbach et al., 2004; Shaw et al., 2012). Additionally, AUC indicated a good general detection accuracy in the Control (unidirectionally $AUC = .76$) and Misdirection condition (bidirectional $AUC = .82$). When differentiating strategy levels, we found that liars using level 1 strategies were well detected by the test (72.7 – 85.7%), but the findings for liars using level 2 strategies were less straightforward. Contrary to our expectation half of the liars who used level 2 strategies were detected. Predictive validity of level 1 strategies is good, but not for level 2 strategies. We suggest the following sources of error that may aid in explaining the error in prediction (for both level 1 and level 2 strategies). First, to execute a strategy the participant needs to recognize the correct answer on each trial. Although memory performance was good, it was not perfect. That means participants either had to guess or selected, from their perspective, an unintended answer on trials for which they did not

remember the correct answer. These errors can easily artificially inflate test scores for level 1 users. Second, we performed our estimates based on strategy levels and not individual strategies. The problem is that not every strategy per level necessarily produces the same test response. Some level 1 strategies (e.g. Demeanor) do not refer to test scores at all, while others do (e.g. 'Avoid correct information'). There are also two potential sources of error for level 2 strategies. First, the concept of level of chance may be hard to grasp. We noted that a considerable proportion (33.33%) of detected liars, who utilized level two strategies, just passed the detection threshold by one answer. In other words, participants may have been unable to correctly determine how many correct answers were necessary to remain within chance levels. Second, we considered the entire test performance as a representation of the reported strategy. However, we are unable to determine the exact moment a strategy was implemented or whether a strategy change took place. Devising or changing to a level 2 strategy during the test may be too late to implement it correctly. Finally, both of these sources of error are further strengthened by the fact that total number of test items was unknown.

Despite this imperfect relationship between strategy level and classification rate, we argue that strategy selection provides a better theoretical construct for behaviour in a forced choice memory performance testing than pure avoidance motivation. So far the latter has simply been assumed, and it can only reasonably explain cases wherein the liar was detected, which is often less than 50% of the data. Strategy selection is supported by the fact that liars in our (and in Shaw et al., 2012) study report using different strategies. These can be conceptualized within a CHT (Carmerer et al., 2004) framework and we also found average strategy levels for liars (Control = 1.59 & Misdirection = 1.41) similar to studies Carmerer et al. (2004) refer to. Although imperfect, strategy selection has at least the same predictive validity as pure avoidance motivation. In addition, it enables predictions for detected and undetected cases and has identifiable sources of error.

Regarding the Misdirection manipulation, by presenting the test situation as a polygraph examination, we were able to reduce the likelihood to realise the test's true mechanism in truth tellers, 30.4% in the Control condition understood the test, but none in the Misdirection condition. This was not the case for liars (Control: 35% and Misdirection: 22%). Additionally, our misdirection led liars (17.4%) and truth tellers (21.7%) to report using physiological countermeasures as their strategy to defeat the test. Although our misdirection did not lower the likelihood for liars to see through the test's mechanism we still found that a considerable proportion reported physiological countermeasures as their strategy. These findings may seem at odds with each other, but a potential explanation could be that only participants fell for the misdirection, who would not have understood the test in the first place.

In terms of test scores we found an increased presence of overperformance in the Misdirection condition. This can be seen in the difference between the uni- and bidirectional criteria, as the latter only improves detection accuracy in the presence of overperformance. In the Control condition we found that both the unidirectional ($AUC = .76$) and the bidirectional ($AUC = .72$) criterion discriminated truth tellers from liars. This was not the case in the Misdirection condition. Here, only the bidirectional criterion ($AUC = .82$) proved better than chance. This suggests that by manipulating the information content the test situation provides test behaviour can be shaped accordingly.

There are two limitations we would like to address. First, in deception detection experiments the mock crime procedure is often criticised for not being realistic enough. We argue that this is not the case here. In forced choice memory testing only one element of a mock crime matters: That it induces the memory of details later encountered in the test. We have measured memorability and consider it high.

Second, we used self reported data to measure the strategies participants used. The validity of self reported data has been subject to discussion (Nisbett & Wilson, 1977; Ericsson, & Simon, 1980), raising the question whether participants can know, in this case, what kind of strategy they actually used. Ericsson & Simon (1980) show that self reported information is reliable if it has been subject to focal attention and at least been in the short term memory. In other words the participant must have been aware of the information to verbalize. Our analyses are based on the strategy levels. This categorization can be reduced to the belief a participant held over the test's mechanism. This information was accessible to participants and therefore can be used.

From a theoretical point of view this study proposes a new perspective on the psychological processes involved in forced choice memory performance testing. We argue that examinees design a strategy to defeat the test and that their strategy selection process can be influenced by managing the information content of the test situation. This study shows that new behaviours can be elicited by drawing on the particular strategy selection process made by examinees, in this case overperformance, through a misdirection of reason.

References

- Bianchini, K.J., Mathias, C.W., & Greve, K.W. (2001). Symptom validity testing: A critical review. *The Clinical Neuropsychologist*, 15(1), 19-45. DOI:10.1076/clin.15.1.19.1907
- Binder, L.M., Larrabee, G.J., & Millis, S.R. (2014). Intent to fail: Significance testing of forced choice test results. *The Clinical Neuropsychologist*, 28(8), 1366 – 1375. DOI: 10.1080/13854046.2014.978383
- Carmerer, C.F., Ho, T., & Chong, J. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3), 861-898. DOI: 10.1162/0033553041502225
- Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups – partial remembering. *Journal of Police Science and Administration*, 1, 287-293.
- Giger, P., Merten, T., Merckelbach, H., & Oswald, M. (2010). Detection of feigned crime-related amnesia: A multi-method approach. *Journal of Forensic Psychology Practice*, 10, 440-463. DOI: 10.1080/15228932.2010.489875
- Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215 – 251.
- Hancock, Peter. Psychological image collection at sterling (PICS). University of Stirling, Stirling. 30 October 2014. <http://pics.psych.stir.ac.uk>
- Jelicic, M., Merckelbach, H., van Bergen, S. (2004). Symptom validity testing of feigned amnesia for a mock crime. *Archives of Clinical Neuropsychology*, 19, 525-531. DOI: 10.1016/j.acn.2003.07.004
- Kuhn, G., Caffaratti, H.A., Teszka, R., & Rensink, R.A. (2014). A psychology-based taxonomy of misdirection. *Frontiers in Psychology*, 5, 1-14. DOI: 10.3389/fpsyg.2014.01392.

- Meijer, E.H., Smulders, F.T., Johnston, J.E., & Merckelbach, H. (2007). Combining skin conductance and forced choice in the detection of concealed information. *Psychophysiology*, 44, 814-822. DOI: 10.1111/j.1469-8986.2007.00543.x
- Merckelbach, H., Hauer, B., & Rassin, E. (2002). Symptom validity testing of feigned dissociative amnesia: A simulation study. *Psychology, Crime, & Law*, 8, 311-318. DOI: 10.1080/1068316021000054256
- Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231 – 259.
- Pankratz, L., Fausti, S.A., & Peed, S. (1975). A forced-choice technique to evaluate deafness in the hysterical or malingering patient. *Journal of Consulting and Clinical Psychology*, 43(3), 421-422. DOI: 10.1037/h0076722
- Shaw, D. J., Vrij, A., Mann, S., Leal, S., & Hillman, J. (2012). The guilty adjustment: Response trends on the symptom validity test. *Legal and Criminological Psychology*. DOI: 10.1111/j.2044-8333.2012.02070.x
- Siegel, S. (1956). *Nonparametric statistics*. New York: McGraw-Hill.
- Van Oorsouw, K., & Merckelbach, H. (2010). Detecting malingered memory problems in the civil and criminal arena. *Legal and Criminological Psychology*, 15, 97 – 114. DOI: 10.1348/135532509X451304
- Verschuere, B., Meijer, E., & Crombez, G. (2008). Symptom validity testing for the detection of simulated amnesia: Not robust to coaching. *Psychology, Crime, & Law*, 14(6), 523-528. DOI: 10.1080/10683160801955183

Table 1.

Self reported strategies distinguished between conditions and strategy levels.

Truth teller	Liar
--------------	------

Strategy	Control	GSR	Control	GSR
<u>Level 0</u>				
No strategy	30.4	39.1	4.3	4.3
<u>Level 1</u>				
Avoid correct information	8.7	-	21.7	30.4
Demeanor / body language	17.4	-	8.7	-
Physiological countermeasures	-	21.4	-	17.4
<u>Level 2</u>				
Imitate ignorance	13	13	30.4	17.4
Deductive guessing	21.7	13	8.7	4.3
Mixture of truth & lie	-	4.3	21.7	21.7
Other	8.7	8.7	4.3	4.3

Notes. Frequency of strategies indicated in percentages.

Table 2.

Number of detected and undetected strategies differentiated by level for Liars in the Control condition (unidirectional) and Misdirection condition (bidirectional).

	Strategy level	N	Detected	Undetected
Control				
	0	1	0	100
	1	7	85.7	14.3
	2	14	50	50
Misdirection				
	0	1	100	0
	1	11	72.7	27.3
	2	10	50	50

Notes. Detection accuracy indicated in percentages.

Table 3.

Frequency of liars' level 2 strategies

Strategy	Detected	Undetected
Imitate ignorance	6	5
Deductive guessing	1	2
Mixture of truth & lie	5	5

Notes. Control (unidirectional) and Misdirection (bidirectional) conditions combined as they had similar distributions of strategies.

Table 4.

Classification Accuracy

	Sensitivity	Specificity	AUC	95% CI
Control				
Unidirectional	56.52	86.95	.76*	.62 - .90
Bidirectional	60.87	86.95	.72*	.56 - .88
Misdirection				
Unidirectional	43.5	100	.67	.49 - .84
Bidirectional	65.22	95.65	.82*	.69 - .95
Notes. * $p < .05$. Cutoff for sensitivity and specificity at $p = .05$				