Concealed Avoidance Testing

Response Strategies of Instructed Malingerers during Forced Choice Testing: New Measures and Criteria to detect Concealed Knowledge and Feigned Cognitive Deficits

Robin Orthey

University of Portsmouth

Department of Psychology

The thesis is submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy of the University of Portsmouth

27th November 2018

Supervised by: Professor Aldert Vrij

Dr Sharon Leal

Dr Hartmut Blank
General Abstract

The Forced Choice Test (FCT) can be used to detect malingered loss of memory for a specific event or malingered cognitive deficits. This test consists out of binary multiple choice questions. The idea is that genuine impairment will force examinees to guess, resulting in test scores that fall within chance performance. In contrast, malingerers are expected to select incorrect answers purposefully, leading to test scores below chance performance (i.e. underperformance). Four empirical studies were conducted on the FCT with two aims in mind. First, to develop a better theoretical understanding of malingerers’ response strategies in the FCT. Second, to increase the detection accuracy of the FCT and optimize the decision making process for single case decision.

In Chapter 2 we address the lack of a theoretical foundation for malingerers’ behaviour by proposing a model that defines three distinct response strategies. With an empirical experiment we demonstrate that the proposed model fits the data well and conclude that that malingerers’ choice of response strategy can be influenced by the examiner. Furthermore, our model suggests that the traditionally used underperformance criterion is only sensitive to one of the three subgroups and that this group is actually a minority within the malingerer population. Based on these results we propose two pathways to improve the detection accuracy of the FCT.

First, detection accuracy can be improved by promoting the prevalence of response strategies the underperformance criterion detects well. In Chapter 3 we attempt to do that by introducing cognitive load to the FCT paradigm. However, instead of affecting malingerers’ strategy selected, cognitive load affected the quality of their chosen strategies. Although unexpected, these findings provide another angle to influence the detection rate of the FCT. Nonetheless, further disambiguation is required.
Second, detection accuracy can be improved by adding new criteria to the FCT that are sensitive to the remaining subgroups. In Chapters 4 and 5 we investigate the effectiveness of the ‘runs test’ and a within test response bias. Both criteria proved effective in detecting a subgroup that manages to avoid the underperformance criterion and the usefulness of a two-step classification procedure is discussed.

Finally, this thesis ends with a reflection on the validity of our model based on the aggregated data from all previous chapters and a discussion about the decision making process. This is followed by a reflection on experimental limitations as well as recommendations for practical application.
Contents

Declaration 9

List of Tables 10

List of Figures 11

Abbreviations 11

Dissemination 12

Chapter 1: General Introduction 13

1.1 Introduction 14

1.1.1 Malingering and the validity of psychological examination 14

1.1.2 The Forced Choice Test 15

1.2 Gap in literature and thesis outline 16

1.3 References 25

Chapter 2: Strategy and Misdirection in Forced Choice memory Performance Testing in Deception Detection 29

2.1 Abstract 30

2.2 Introduction 31

2.3 Method 36
Chapter 3: Effects of Time Pressure on Strategy Selection and Strategy Execution in Forced Choice Testing

3.1 Abstract

3.2 Introduction

3.3 Method

3.3.1 Participants

3.3.2 Procedure

3.3.3 Materials

3.3.4 Design

3.4 Results

3.4.1 Understanding and misdirection

3.4.2 Strategies

3.4.3 Avoidance behaviour and detection accuracy

3.5 Discussion

3.6 References
3.4 Results

3.4.1 Reaction time

3.4.2 Strategy levels

3.4.3 Test scores

3.5 Discussion

3.6 References

Chapter 4: Resistance to Coaching in Forced-Choice Testing

4.1. Abstract

4.2. Introduction

4.3 Method

4.3.1 Participants

4.3.2 Procedure

4.3.3 The Forced-Choice Test

4.3.4 Design and measures

4.4. Results

4.4.1 Strategies

4.4.2 Detection accuracy

4.4.3 Incremental validity

4.5 Discussion
Chapter 5: Eliciting Response Bias Within Forced Choice Tests to Detect Random Responders

5.1 Abstract

5.2 Introduction

5.3 Method

5.3.1 Participants

5.3.2 Procedure

5.3.3 Design

5.4 Results

5.5 Discussion

5.6 References

Chapter 6: General Discussion

6.1 General discussion

6.2 The Three Strategy Levels and Cognitive Hierarchy Theory

6.3 Detection accuracy

6.4 How to improve Detection Accuracy

6.5 Limitations

6.6 Test Construction and Practical Application
6.7 Innovation in Practical Application 145

6.8 Challenges and Future Directions 146

6.9 Conclusions 148

6.10 References 150

Appendices 155

Appendix A: Favourable Ethical Opinion (Chapter 2) 155

Appendix B: Favourable Ethical Opinion (Chapter 3) 156

Appendix C: Favourable Ethical Opinion (Chapter 4) 159

Appendix D: Favourable Ethical Opinion (Chapter 5) 160

Appendix E: UPR16 Form 161
Declaration

Whilst registered as a candidate for the above degree, I have not been registered for any other research awards. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

___________________________________
Robin Orthey

WORD COUNT = 33.126 words

Word count after revisions = 34.001
### List of Tables

1.1 Overview of empirical studies on the Forced Choice Test ........................................ 18

2.1 Self reported strategies distinguished between conditions and strategy levels .................. 44

2.2 Number of detected and undetected strategies differentiated by level for Liars in the Control condition (unidirectional) and Misdirection condition (bidirectional) ............... 46

2.3 Frequency of liars’ level 2 strategies .............................................................................. 47

2.4 Classification accuracy .................................................................................................... 48

3.1 Detection accuracy of total scores per condition ............................................................. 66

3.2 Percentage of total scores per strategy level categorized into Under-, Chance-, and Overperformance for examinees with concealed knowledge ......................................... 67

4.1 Frequencies of strategy levels per condition ................................................................ 89

4.2 Detection accuracy for the alternation criterion ............................................................. 92

4.3 Detection accuracy of two-step classification using total score criterion and the number of runs criterion .............................................................................................................. 94

5.1 Detection accuracies for all criteria for the Standard and Opacity condition ............... 116

6.1 Malingers’ average test z-scores of number of correct answers selected combined and separated per strategy level .................................................................................. 133
List of Figures

4.1 Receiver operating characteristic curve for correct total and alternation criteria for naïve and coaching conditions 93

5.1 AUCs of the Runs test in the Standard and Opacity condition as well as the Response bias in the Opacity condition for all test lengths from 12 to 100 trials 117

6.1 Histogram of malingerers’ z-transformed test scores per strategy level 134

6.2 Distribution of z-scores for the empirical (top) and simulated (bottom) samples of malingered and genuine test performance 136

Abbreviations

FCT Forced Choice Test / Forced Choice Testing

CHT Cognitive Hierarchy Theory

SVT Symptom Validity Test

SCR Skin Conductance Response

ROC Receiver Operating Characteristic

AUC Area Under the Curve
Dissemination

Publications


Chapter 1:

General introduction
1.1 Introduction

1.1.1 Malingering and the validity of psychological examination

Malingering, the simulation or exaggeration of cognitive deficits or memory loss, is a serious concern for the interpretation of psychological examinations. Slick (1999) suggested that malingering should be considered a possibility when a clinical diagnosis yields an incentive, such as a financial compensation or reduced sentencing in criminal cases. For example, around 25% of homicide cases feature claims of partial or complete memory loss (Cima, Nijman, Merckelbach, Kremer, & Hollnack, 2004). As the loss of memory may aid a suspect during his/her trial, the suspect may be motivated to feign the symptoms. The prevalence of malingering varies, but a general estimate suggests that between 20 to 30% of civil and criminal cases may be cases of malingering (see Cima et al., 2004; Mittenberg, Patton, Canyock, & Condit, 2002).

The consequences of malingering can be severe (see Rosen, & Phillips, 2004). First, Chafnetz and Underhill (2013) showed that feigned disability can incur large monetary costs due to fraudulent disability compensation. Second, malingering can cast doubt about the validity of a psychological examination. Consequently, institutions such as the American Academy of Clinical Neuropsychology (AACN; Chafetz et al., 2015; Heilbronner, Sweet, Morgan, Larrabee, Millis, & conference participants, 2009) and the Association for Scientific Advancement in Psychological Injury and Law (ASAPIL; Bush, Heilbronner, & Ruff, 2014) propose that malingering assessment tools should be included in a standard psychological examination.
1.1.2 The Forced Choice Test

The Forced Choice Test (FCT) can be used as a tool to detect malingered sensory impairment (Pankratz, 1983). In essence, the FCT includes several trials, each presented with two possible answer alternatives. The examinee is asked to indicate the correct answer alternative, or to guess in case s/he is unaware. The total number of correct items selected is then interpreted in terms of how (un)likely it is to occur through guessing alone. For example, Pankratz, Fausti, and Peed (1975) carried out a FCT on a patient who claimed to suffer from loss of hearing. The patient was placed in front of a light that could either turn red or blue for two seconds. During one of the two trials, determined randomly, a 1000Hz tone was presented and the patient had to indicate on which trial (red or blue) he heard the tone. This procedure was repeated 100 times and the patient identified the correct trial 36 times. Genuine loss of hearing would force the patient to guess on each trial, resulting in approximately 50 correct responses selected. Consequently, the observed test score can be expressed as the likelihood of occurring under chance performance according to the binomial distribution (see Siegel, 1956). In this case, 36 correct trials out of 100 leads to a p-value smaller than .004, which means that if the test was repeated 1000 times a patient with genuine impairment would be expected to produce less than four test results with a score like this or more extreme. Consequently, Pankratz, Fausti, and Peed (1975) concluded that the total score was too unlikely to have occurred by chance and, instead, was the product of deliberately selecting incorrect answers.

Similarly, the FCT can be used to detect deceit in terms of malingered loss of memory for specific events. Denney (1996) was asked to assess the competency of three suspects to stand trial in three criminal cases. All three suspects claimed to have suffered memory loss during the offense. A unique FCT was created for each suspect specific to the offense. In each case the suspect was presented with a question (e.g., Was the gun fired?)
and two equally plausible answer alternatives (Yes/No). Suspects were instructed to select the correct answer alternative or guess in case they did not know it. All three suspects produced a total score very unlikely to occur by chance with fewer correct items selected than expected by chance performance, suggesting they purposefully selected incorrect answers to demonstrate their memory loss.

1.2 Gap in literature and thesis outline

Table 1.1 presents an overview of recent experiments investigating the diagnostic accuracy of the FCT in detecting malingered loss of memory for an event, excluding case studies. Together, these studies suggest a detection rate for malingerers – sensitivity – between 40 – 60% and a detection rate for genuine impairment – specificity – around 95% (see Giger, Merten, Merckelbach, & Oswald, 2010; Jelicic, Merckelbach, & van Bergen, 2004; Meijer, Smulders, Johnston, & Merckelbach, 2007; Merckelbach, Hauer, & Rassin, 2002; Shaw, Vrij, Mann, Leal, & Hillman, 2012; Verschuere, Meijer, & Crombez, 2008). When warned, sensitivity of the FCT declines significantly (see Giger et al., 2010; Verschuere et al., 2008). It is noteworthy, that these accuracy estimates are predominantly based on a single cut-off alone. Only two studies (Meijer et al., 2007; Shaw et al., 2012) report the general detection accuracy in terms of the Area Under the Curve (AUC), which is a sum of the detection accuracy estimates for all possible cut-offs. Finally, several studies lack control groups that is, specificity estimates, and many feature only minimal sample sizes. With these limitations and given the small number of experiments in total, the detection accuracy of the FCT remains far from certain. Hence, further research in the diagnostic utility of the FCT is needed.
Cases of underperformance can be attributed to intentional avoidance of correct answers (Pankratz, Fausti, & Peed, 1975). However, considering the estimated sensitivity, this explanation typically accounts for only half or less of the malingerers. Furthermore, when asked, some malingerers indicated understanding of the FCT rationale and also report other behaviours than strictly avoiding correct answers (Jelicic et al., 2004; Shaw et al., 2012). Currently, malingerers’ response strategies in the FCT lack a proper conceptualization. As such, it remains unclear what the specific strategies are and how sensitive the underperformance criterion is to each of them. Similarly, it is unknown how malingerers’ decide what response strategy to follow. Further investigation into malingerers’ response strategies is warranted, because this knowledge can serve as the basis to improve the detection accuracy further.

The objective of this thesis is to develop a model of malingerers’ behaviour during a FCT. This thesis proposes a model that specifies three distinct response strategies and to suggest to use this model to modify the FCT paradigm to increase the overall detection accuracy. This PhD thesis features four studies reported in four separate chapters. These chapters are based on peer reviewed articles or articles submitted for peer review. They will contain some repeated information - each published article must be understandable on its’ own-and each chapter will feature its own reference section.
<table>
<thead>
<tr>
<th>Study</th>
<th>Critical (Total) FCT items</th>
<th>Sensitivity (N)</th>
<th>Specificity (N)</th>
<th>AUC</th>
<th>Under-performance</th>
<th>Chance performance</th>
<th>Over-performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merckelbach et al., 2002</td>
<td>15 (15)</td>
<td>40% (20)</td>
<td>-</td>
<td>-</td>
<td>40% (8)</td>
<td>30% (6)</td>
<td>30% (6)</td>
</tr>
<tr>
<td>Jelicic et al., 2004</td>
<td>25 (50)</td>
<td>59% (39)</td>
<td>-</td>
<td>-</td>
<td>59% (23)</td>
<td>26% (10)</td>
<td>15% (6)</td>
</tr>
<tr>
<td>Van Oorsouw &amp; Merckelbach, 2006*</td>
<td>21 (40)</td>
<td>7% (27)</td>
<td>100% (30)</td>
<td>-</td>
<td>7% (2)</td>
<td>60% (16)</td>
<td>33% (9)</td>
</tr>
<tr>
<td>Meijer et al., 2007 – Study 1</td>
<td>12 (12)</td>
<td>27% (30)</td>
<td>100% (30)</td>
<td>.70</td>
<td>27% (8)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Meijer et al., 2007 – Study 2</td>
<td>12 (12)</td>
<td>- (60)</td>
<td>- (60)</td>
<td>.87</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Verschuere et al., 2008 – Naïve</td>
<td>25 (35)</td>
<td>58% (19)</td>
<td>-</td>
<td>-</td>
<td>58% (11)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Verschuere et al., 2008 – Coached</td>
<td>25 (35)</td>
<td>0% (19)</td>
<td>-</td>
<td>-</td>
<td>0% (0)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Giger et al., 2010 – Naïve</td>
<td>19 (38)</td>
<td>45% (20)</td>
<td>90% (20)**</td>
<td>-</td>
<td>45% (9)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Giger et al., 2010 – Warned</td>
<td>19 (38)</td>
<td>10% (20)</td>
<td>90% (20)**</td>
<td>-</td>
<td>10% (2)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Shaw et al., 2014*</td>
<td>12 (12)</td>
<td>42% (86)</td>
<td>93% (82)</td>
<td>.79</td>
<td>42% (36)</td>
<td>57% (49)</td>
<td>1% (1)</td>
</tr>
</tbody>
</table>

Notes. Studies marked with a * should be interpreted with care as they suffer methodological concerns such as biased FCT questions or errors in cut-off calculation. Specificity marked with ** refers to the same group. Giger et al., 2010 featured three groups only: naïve malingierer, warned malingierer, and examinees without crime knowledge (marked with **).
Throughout the next four chapters the experiments will feature cases of malingered loss of memory for crime events or malingered colour blindness. Consequently, different terms may be used for malingerers and genuine performance. Malingerers, examinees instructed to simulate loss of memory or sensory deficits, may be referred to as ‘liars’, ‘examinees with concealed knowledge’, or ‘malingerers’. Genuine performance, through real impairment or ignorance of crime information, may be referred to as ‘truth tellers’, ‘examinees without concealed knowledge’, or ‘genuine performance’.

First, a new model for malingerers’ response strategies in the FCT is developed (chapter 2). Based on this model two approaches to increase detection accuracy of the FCT are investigated. On the one hand, detection accuracy could be increased by facilitating malingerers to choose response strategies that the FCT already detects well (chapter 3). On the other hand, detection accuracy could be increased by examining the validity of additional criteria that are sensitive to previously undetected subgroups of malingerers (chapters 4 and 5).

Chapter 2: Strategy and Misdirection in Forced Choice Memory performance Testing in Deception Detection

This chapter deals with the lack of theoretical conceptualization of malingerers’ behaviour during a FCT. To address this, a model was devised based on Cognitive Hierarchy Theory (CHT; Carmerer, Ho, & Chong, 2004) and strategies derived from self-reports reported in previous studies. According to CHT, examinees take the strategies of other players into account when developing their own strategy and the sophistication of this process is limited by the available
cognitive resources of the examinee. Another feature is that strategies are hierarchical, indicated with numeric levels, with higher levels being superior to lower. Here three strategy levels for malingers in the FCT ranging from 0 to 2 were defined. A level 0 strategy predicts the examinee will endorse the correct information, resulting in overperformance (more items correct than expected by chance). A level 1 strategy predicts that the examinee will avoid selecting correct answers, resulting in underperformance. A level 2 strategy predicts that an examinee will provide a balanced mixture of correct and incorrect answers. Liars and truth tellers were subjected to a traditional FCT or to a FCT that included a fake polygraph examination to misdirect examinees’ beliefs about the detection mechanism of the FCT. Test performance and response strategy levels were measured. The main findings were that (i) substantial proportions of liars used level 2 strategies, which suggests correct understanding of the FCT’s detection mechanism; (ii) different strategy levels featured different detection accuracies; and (iii) examinees test behaviour could be influenced by misdirecting them from the test detection mechanism. Together, these findings provide the first support for our proposed model. Based on this knowledge, future manipulations of the FCT paradigm can be developed to increase the detection accuracy of the FCT. This experiment has been published in Orthey, Vrij, Leal, and Blank (2017).

Chapter 3: Effects of Time Pressure on Strategy Selection and Strategy Execution in Forced Choice Testing

This chapter draws on the model introduced in the previous chapter. As stated, the FCT is apt at detecting level 1 strategies, but not level 2 strategies. The
underlying theory assumes that higher order strategies require more cognitive resources. Therefore, if cognitive resources are limited examinees may be less likely to select a strategy level with poor detection accuracy (level 2) and instead promotes selection of lower order strategies (level 0 and 1), which are more easily detected. To limit the available cognitive resources time pressure was introduced to the FCT paradigm, by forcing examinees with and without concealed information to select an answer alternative within two seconds. This paradigm was compared with the traditional FCT in terms of strategy selection and detection accuracy. The main findings were that (i) selection of strategy levels was not affected by time pressure; (ii) in both the time pressure and traditional FCT, the number of correct items selected discriminated examinees with and without concealed knowledge better than chance; and (iii) examinees with concealed knowledge, who reported a level 2 strategy, had selected fewer correct items under time pressure than in the traditional FCT paradigm, leading to a considerable higher proportion of cases at underperformance level in the time pressure FCT. These results suggest that time pressure is not suited to affect the strategy selection process, but instead affects execution of the strategy. That is, examinees who report level 2 strategies and intend to randomize between correct and incorrect answers, are expected to avoid detection by the underperformance criterion, as demonstrated in the standard condition limiting the overall detection accuracy of the FCT. The time pressure condition demonstrated that examinees, using the same strategy, were less successful in avoiding detection by the underperformance criterion. Consequently, cognitive load, in terms of time pressure, could be used to limit the effectiveness of a common and effective counterstrategy in the FCT.
Chapter 4: Resistance to Coaching in Forced-Choice Testing

This chapter examinees a new criterion to detect level 2 strategies and its value in dealing with cases of coaching. Coaching describes the act of an examinee seeking information on a forensic test prior to administration. This is a concern for the FCT, because once an examinee is aware of the underperformance criterion the examinee is likely to use a level 2 strategy and randomize between correct and incorrect answers. As the detection rate for level 2 strategies is poor, coaching is a threat to the validity of the FCT. The ‘runs-test’ has been suggested to measure the alternations between correct and incorrect answers. It is based on the idea that examinees who are unaware of the correct answer, have a likelihood of 50% to alternate between trials, thus like the traditional criterion, they are expected to produce a number of alternations within chance levels. In contrast, examinees who are aware of the underperformance criterion, are expected to alternate more frequently between correct and incorrect answers to ensure that the total number of correct items falls within chance levels. Hence, the ‘runs-test’ detects examinees using a level 2 strategy through elevated alternation rates between correct and incorrect answers. So far, empirical support suggests it is of limited value only (Jelicic et al., 2004; Verschuere et al., 2008). To increase the validity of the runs-test, examinees were forced to choose between a randomizing pattern that ‘looks’ random and a randomizing pattern that produces a test score within chance levels. To do so the position of the correct answer alternative (left or right) was alternated between trials as well. Consequently, alternating between correct and incorrect answers means only answers on the same side are selected, while alternating between answers presented on the left and right would lead to more extreme scores. Coached examinees were expected to prefer the former, while examinees who are genuinely
guessing, would prefer the latter pattern. Detection accuracy of the ‘runs-test’ would be increased, because both types of randomizing behaviour are anti-correlated, increasing the difference between both groups. The main findings were that (i) coaching was associated with level 2 strategies and underperformance was apt at detecting level 1, but not level 2 strategies; (ii) the runs-test was able to detect coached examinees with concealed knowledge; (iii) the underperformance criterion and runs-test criterion can be utilized as a 2-step classification procedure, with underperformance being sensitive to level 1 strategies and the runs-test to level 2 strategies. Together these findings support the underlying strategy levels and their associated test scores as well as the need to detect level 2 strategies in order to increase detection accuracy of the FCT. This article was published as Orthey, Vrij, Meijer, Leal, and Blank (2018).

Chapter 5: Eliciting Response Bias Within Forced Choice Tests to Detect Random Responders

This chapter is focused on detecting level 2 strategies in case of malingered red/green blindness. Specifically, the validity of the ‘runs-test’ and a new criterion was evaluated. The idea behind the new criterion was to elicit a response bias within the generation process of a test score that falls within chance performance. Specifically, the aim was to introduce a manipulation, independent of the actual task (here to discriminate red and green), that elicits a systematic preference. The perceived difficulty of the trials was varied by manipulating the see-throughness of the stimuli, so malingerers could be more likely to select correct answers on trials that are clearly visible and be more likely to select incorrect answers when they are
not. If examinees attune their selection preference to the perceived difficulty of the trial, this systematic pattern would deviate from genuine guessing behaviour and could serve as a new criterion. Therefore, examinees were instructed to simulate red/green blindness and subjected them to a FCT of 100 trials embedded into a test battery. The FCT in the standard condition featured a bright red and a bright green rectangle on each trial. Additionally, in the opacity condition, the see-throughness of the rectangles was varied over the trials, creating the illusion that on some trials the correct answer would be more difficult to identify than on others. The validity of the ‘runs-test’ was re-examined, because malingered sensory deficits, as opposed to malingered loss of memory, allows for the construction of FCTs with larger test sizes and the ‘runs-test’ should be more effective with longer tests. The main findings were that: (i) the runs-test did detect malingerers better than chance in the standard, but not in the opacity condition; (ii) malingerers produced more statistically significant response biases than expected by chance; and (iii) the probability of the individual response biases detected malingerers better than chance.

These findings suggest that the ‘runs-test’ or a systematic association between perceived trial difficulty and endorsement of correct answers could be used as criterion to detect level 2 strategies in malingered sensory deficits.

**Chapter 6: General Discussion**

Finally, the concluding chapter will re-examine the theoretical model suggested by the previous studies, discuss how detection accuracy should be estimated in a FCT, and evaluate the previous studies in terms of their practical relevance.
1.3 References


Chapter 2:

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

This chapter is based on:

2.1 Abstract

We examined the response strategies of liars in the Forced Choice Test (FCT). Various response strategies have been observed and we attempt to categorise these strategies within the framework of Cognitive Hierarchy Theory. Furthermore, we investigated whether liars’ response strategies could be shaped through misdirection. 95 undergraduate students were subjected to either a mock crime or a filler task. Then participants were either subjected to a standard FCT or a FCT that additionally featured a fake polygraph setup to distract from the FCT’s test mechanism. The results suggest that liars’ response strategies fit one of three categories and the categories correspond well to the observed test scores. The misdirection manipulation lowered the likelihood to see through the test’s mechanism for truth tellers, but not for liars. Finally, liars’ response patterns were shifted by the misdirection manipulation. Together these results support our proposed categories of liars’ response strategies.

1 This abstract differs from the published version.
2.2 Introduction

Concealed information detection is an indirect deception detection approach. The idea is to detect knowledge in suspects that only the investigators and the perpetrator are aware of and 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 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; Jelicic, 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.

Two studies provide hints to the underlying mechanism of forced choice memory performance testing. First, 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. Second, Verschuere, 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. Together 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. Carmerer 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.

2.3 Method

2.3.1 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.

2.3.2 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.

2.3.3 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)’. The first question was used to determine the strategy each participant used. It was directed at the participants actual behaviour instead of
conception of strategy to avoid biases introduced by the question, see Schwarz (1999). The latter was used 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.

2.3.4 Design

This study 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

---

2 The published version claimed to have used a double blind design. This was not the case.
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).

2.4 Results
2.4.1 Understanding and 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.

2.4.2 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 2.1 Self reported strategies distinguished between conditions and strategy levels.
Next, we examined the strategy levels. We conducted a Chi-square test to compare the distributions of strategy levels for truth tellers and liars for the standard and misdirection condition. In the standard condition liars were more likely to exhibit a higher level strategy than truth tellers, $X^2(2, N = 43) = 6.19, p = .045$.

Similarly, in the misdirection condition liars were also more likely to exhibit a higher level strategy than truth tellers, $X^2(2, N = 46) = 9.16, p = .010$. These findings support H2. In addition we looked at the actual prevalence of strategy levels for each

---

3 These analyses differ from the published version. Originally, we conducted a 2x2 ANOVA and a number of one-sample t-tests. The conclusions drawn did not change.
group. Liars in the standard condition (level 0 = 4.5%; level 1 = 31.8%; level 2 = 63.6%) and misdirection condition (level 0 = 4.5%; level 1 = 50%; level 2 = 45.5%) display primarily level 1 and level 2 strategies. Level 0 strategies occur only rarely. Truth tellers in the standard condition (level 0 = 33.3%; level 1 = 28.6%; level 2 = 38.1%) and misdirection condition (level 0 = 42.9%; level 1 = 23.8%; level 2 = 33.3%) displayed much higher frequencies of level 0 strategies, but in contrast to our expectations they also displayed level 1 and level 2 strategies. In fact the we found that only less than half of our truth tellers followed a level 0 strategy. 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 2.2 Number of detected and undetected strategies differentiated by level for Liars in the Control condition (unidirectional) and Misdirection condition (bidirectional).
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 2.3 Frequency of liars’ level 2 strategies

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Detected</th>
<th>Undetected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes. Detection accuracy indicated in percentages.
<table>
<thead>
<tr>
<th>Imitate ignorance</th>
<th>6</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deductive guessing</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Mixture of truth &amp; lie</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

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

2.4.3 Avoidance behaviour and 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 2.4 Classification Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unidirectional</td>
<td>56.52</td>
<td>86.95</td>
<td>.76*</td>
<td>.62 - .90</td>
</tr>
<tr>
<td>Bidirectional</td>
<td>60.87</td>
<td>86.95</td>
<td>.72*</td>
<td>.56 - .88</td>
</tr>
<tr>
<td>Misdirection</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unidirectional</td>
<td>43.5</td>
<td>100</td>
<td>.67</td>
<td>.49 - .84</td>
</tr>
<tr>
<td>Bidirectional</td>
<td>65.22</td>
<td>95.65</td>
<td>.82*</td>
<td>.69 - .95</td>
</tr>
</tbody>
</table>

Notes. * p < .05. Cutoff for sensitivity and specificity at p = .05

2.5 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.
Relative to previous studies (Giger et al., 2010; Jelicic et al., 2004; Meijer et al., 2007; Merckelbach et al., 2002; & Shaw et al., 2012) we found a high detection accuracy for liars (56.20 – 65.22%) and slightly lower than average (86.95%) to excellent (95.65%) specificity. Our specificity falls within the range of previous studies (Giger et al., 2010; Meijer et al., 2007; & Shaw et al., 2012). Small fluctuations are to be expected, as these groups represent actual chance performance, which means that a priori defined specificities will be approached with increases of sample size. 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.
2.6 References


Chapter 3:

Effects of Time Pressure on Strategy Selection and Strategy Execution in Forced Choice Testing

This chapter is submitted as:

3.1 Abstract

We investigated the effects of cognitive load on the detection accuracy and development of counter strategies of examinees with concealed knowledge in the Forced Choice Test (FCT). Specifically, we hypothesized that cognitive load would force participants to develop less effective counter strategies, leading to an increase in detection accuracy. We subjected 120 participants with or without concealed knowledge to either a standard FCT or a FCT under cognitive load. Cognitive load was induced through time pressure. In both conditions examinees with concealed knowledge were detected better than chance and the best detection accuracy was found in the cognitive load condition. Furthermore, cognitive load did not lower the incident rate of effective counter strategies, but a closer look at the data suggests that instead their success rates were reduced. Further disambiguation of the effects of cognitive load and disentanglement between strategy selection and execution is needed.
3.2 Introduction

The Forced Choice Test (FCT) can be applied to detect concealed knowledge about an event (Denney, 1996; Pankratz, 1983). In a FCT, the examinees are presented with questions about the event, two possible answer alternatives, and the instruction to select the correct answer alternatives or to guess in case they don’t know. While examinees who are unaware of the correct answer have no choice other than to guess, examinees who try to conceal their knowledge tend to purposefully select incorrect answers. Therefore, test scores fall below chance levels – so called underperformance – and can be used as detection criterion (Bianchini, Mathias, & Greve, 2001; Van Oorsouw, & Merckelbach, 2010).

Empirical research suggests that examinees with concealed knowledge can successfully be detected at rates varying from 40% to 60% (Giger, Merten, Merckelbach, & Oswald, 2010; Jelicic, Merckelbach, & van Bergen, 2004; Meijer, Smulders, Johnston, & Merckelbach, 2007; Merckelbach, Hauer, & Rassin, 2002; Orthey, Vrij, Leal, & Blank, 2017; Shaw, Vrij, Mann, Leal, & Hillman, 2012). This detection accuracy is directly related to the prevalence of three different self-reported response patterns that examinees with concealed knowledge use to avoid being detected by the test (Orthey et al., 2017; Orthey, Vrij, Meijer, Leal, & Blank, 2018). These response patterns are defined in terms of hierarchical strategy levels and specify how answer alternatives are selected depending on the examinees beliefs about the test’s detection mechanism (Orthey et al., 2017; Orthey et al., 2018). Examinees using level 0 strategies form no belief about the test’s detection mechanism and comply with the test instructions to select the correct answer alternatives. Examinees using level 1 strategies assume the test’s detection mechanism is based on a level 0 strategy and their response pattern is a reaction to
the test instructions. Instead of selecting the correct answers, examinees select the incorrect answers. Examinees using level 2 strategies assume the test uses a level 1 strategy as detection mechanism and provide a mixture of correct and incorrect answers as response pattern instead. Although, each strategy level predicts different behaviour the intended objective is the same, namely to avoid detection by the FCT. In a FCT, levels 1 and 2 are the most prevalent strategy levels with roughly equal frequencies; level 0 strategies rarely occur in examinees with concealed knowledge. Consequently, the underperformance criterion used to detect concealed knowledge in a FCT is apt at detecting level 1 strategies, but does not detect level 2 strategies. Therefore, in theory, detection accuracy could be increased by manipulations that shift the participant’s strategy from level 2 to level 1.

The three strategy levels were derived from Cognitive Hierarchy Theory (CHT; see Carmerer, Ho, & Chong, 2004). From this theory it follows that limitations in cognitive resources affect the strategy selection. As such, the strategy an examinee selects is not necessarily the optimal strategy, but, rather a strategy that is ‘good enough’ given the available cognitive resources (also known as satisficing; see Simon, 1955). Previous research indicates that a large proportion of examinees have sufficient cognitive resources available to discern the test’s mechanism and to devise an appropriate counter strategy (see Orthey et al., 2017; Orthey et al., 2018). Thus, if one could limit the cognitive resources available to examinees, this would reduce the frequency of higher order strategies (e.g. level 2). As a consequence the detection accuracy of the FCT would increase, because more examinees would be forced to employ a level 1 strategy instead.

It is generally accepted that humans have a limited amount of cognitive resources available at any given moment. Therefore, increasing cognitive load limits
these available resources (Plass, Moreno, & Brunken, 2010). We chose to implement
cognitive load through time pressure, as it is a commonly used manipulation for
cognitive load (see Klapproth, 2008) and it can easily be introduced into the FCT
paradigm. Hence, we subjected examinees to a mock crime procedure or a filler task
followed by either a standard FCT or a FCT with the restriction that each question
has to be answered within two seconds. We tested two hypotheses: Under time
pressure, examinees will be more likely to report using lower level strategies (e.g.
level 1 instead of level 2, or level 0 instead of level 1) than under standard conditions
(Hypothesis 1). As a consequence, examinees with concealed knowledge will display
more extreme (positive or negative) test scores, resulting in increased classification
accuracy of the FCT (Hypothesis 2).

3.3 Method

3.3.1 Participants

We tested 120 participants (33 males, 87 females) from the university
undergraduate population of Bergamo University. Their mean age was $M = 24.61$
($SD = 7.31$). Ethical approval was obtained.

3.3.2 Procedure

Examinees were randomly assigned to one of two Virtual Reality scenarios. In both
scenarios examinees were placed in a virtual apartment that could be freely
explored from the first person perspective. In the concealed knowledge conditions
examinees were told that they were to investigate the apartment of a terrorist and had
to obtain as much information as possible about the terrorist and his planned actions. The apartment contained clues that could be investigated further. These clues were easily visible. Examinees could examine them further by clicking on them. This provided them with a more detailed picture and short description of the clue. Once all clues were examined the simulation terminated and examinees were instructed not to reveal the knowledge gained from the simulation for the remainder of the experiment. In the no concealed knowledge condition, examinees were instructed to survey a different apartment and instructed to remember as much details as possible. This simulation terminated after three minutes.

Then, all examinees were subjected to a FCT examination about the terrorist apartment. The test was computerized and examinees were randomly assigned to either the standard or time pressure condition. In the standard conditions, examinees received 20 questions about the terrorist apartment and each question featured two possible answer alternatives. Questions were presented in two steps. First the question was displayed in the centre of the screen. Upon clicking the ‘next’ button at the bottom centre of the screen the question disappeared and the two answer alternatives were presented at the top left/right side. All answer alternatives were pictures and examinees could select an answer by clicking on it with a mouse button. Examinees received the following instructions: ‘Next, you will be presented with a number of questions and two answer alternatives per question. Select the correct answer. If you don’t know, guess.’ Examinees in the time pressure condition received the additional instruction: ‘You have to choose an answer alternative for each question within two seconds, otherwise the trial will time-out. If you time-out too often you fail the test automatically. In case an examinee took longer than two seconds a buzzer sound occurred to signal the time-out.
After the FCT procedure, examinees were instructed that the deception detection task was over and that they should answer the following questions honestly. They were asked: ‘What did you do to avoid being classified as a liar by the previous test?’ Their answers were recorded, transcribed, and coded by two independent coders.

Finally, examinees with concealed knowledge received the 20 FCT questions and answer alternatives again and were tasked to indicate the correct answer alternative they remembered from the simulation. This served as a memory check and memory performance was good (91.17%).

3.3.3 Materials

We used the same Virtual Reality simulations, FCT questions and answer alternatives as in Orthey et al. (2018). The answer alternatives of all questions were validated to be equally plausible (see Doob & Kirschenbaum, 1973). In total the FCT contained 20 questions with 2 answer alternatives each. Answer alternatives were presented pictorially and had the same size. To control for order effects, the sequence of questions was counterbalanced across examinees, using a latin square of the size 20. Therefore, the 20 questions occurred equally often over all possible trials (1 – 20). The horizontal alignment of the correct answer alternative was determined randomly on each trial.
3.3.4 Design

This experiment featured a 2 Veracity (concealed knowledge vs no concealed knowledge) x 2 Cognitive load (standard vs time pressure) between-subjects design with the test scores as dependent variable. Test scores were computed by submitting the raw total number of correct answer alternatives selected to a z-transformation according to the binomial distribution (see Siegel, & Castellan, 1988, p. 43). The higher/smaller a z-score was, the less likely it was to occur due to chance and smaller scores were indicative of underperformance. Detection accuracy was estimated using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC; see Tanner, & Swets, 1954; Hanley & McNeil, 1982). An ROC plots the sensitivity (detection rate of the signal) against specificity (detection rate for noise) for each possible cut off. The AUC indicates the general diagnostic value for all possible cut-offs. An AUC ranges from 0 to 1 and 0.5 refers to chance performance. AUCs larger than 0.5 suggest that the criterion detects the signal better than chance. In addition, we categorized FCT z-scores into under-, chance-, and overperformance. We handled the traditional 5% cut-off (bidirectional; z-scores larger than 1.96 for overperformance, or lower than -1.96 for underperformance) for classification.

Examinees reaction time was measured in two ways. On each trial reaction time was measured from the time the two answer alternatives were presented up to the time participants selected an answer alternative. First, we computed the average reaction time over all trials per participants. Second, we counted the number of times participants exceeded the two second threshold in the time pressure condition. These measures were used as a manipulation check. Examinees’ responses to the open ended question were coded into distinct strategy levels (0, 1, and 2) (see Orthey et al., 2017; Orthey et al., 2018). These strategy levels define the selection strategy.
of the examinee based on their belief over the tests’ detection mechanism. Level 0 strategies form no belief over the detection mechanism and comply with the test instructions to select the correct answer alternatives. Hence, level 0 strategies would result in overperformance. Level 1 strategies operate on the belief that the test identifies concealed knowledge through complying with the test instructions and therefore, feature a reaction to them, such as picking the incorrect answers instead. Employing level 1 strategies would result in underperformance. Finally, level 2 strategies follow from the understanding that the test detects concealed knowledge through underperformance. Consequently, level 2 strategies feature behaviours with the goal to provide a mixture of correct and incorrect answers, resulting in test scores that fall within chance performance.

3.4 Results

3.4.1. Reaction time

A 2 Veracity (concealed knowledge vs no concealed knowledge) x 2 Cognitive Load (standard vs time pressure) between subjects ANOVA with average reaction time as dependent variable was conducted to assess the effectiveness of our time pressure manipulation. We found a significant difference for Veracity, $F(1,116) = 18.91, p < .001, \eta^2 = .14$. Examinees without concealed knowledge ($mean = 4.99, SD = 4.58$) took longer than those with concealed knowledge ($mean = 2.92, SD = 2.43$). We also found a significant effect for cognitive load, $F(1,116) = 108.92, p < .001, \eta^2 = .48$. Examinees in the standard condition ($mean = 6.43, SD = 3.99$) took longer than examinees in the time pressure condition ($mean = 1.48, SD = 0.81$). There was also a significant Veracity x Cognitive load interaction, $F(1, 116) = 11.11,$
$p = .001$, $\eta^2 = .09$. The difference between examinees with and without concealed knowledge was larger in the standard condition (Concealed knowledge: $mean = 4.61$, $SD = 2.47$; No concealed knowledge: $mean = 8.25$, $SD = 4.42$) than in the time pressure condition (Concealed knowledge: $mean = 1.25$, $SD = 0.22$; No concealed knowledge: $mean = 1.73$, $SD = 1.09$). Furthermore, we looked into the number of times participants timed out (i.e. a response time longer than two seconds) in the time pressure condition. Examinees with concealed knowledge timed out on average 1.03 ($SD = 0.99$) times, and examinees without concealed knowledge timed out on average 4.07 ($SD = 4.52$) times. In sum, we conclude that our time pressure manipulation was effective.

3.4.2 Strategy levels

First, we examined the strategy levels of examinees with concealed knowledge. In both the standard and time pressure conditions, level 1 strategies were the most prevalent (Standard = 48%; Time pressure = 62%) followed by level 2 strategies (Standard = 33%; Time pressure = 28%) and level 0 strategies (Standard = 19%; Time pressure = 10%). A Chi-square test of independence was calculated comparing the frequency of the strategy levels between standard and time pressure condition. We found that the frequency of the different strategy levels did not differ between conditions, $X^2 (2, N = 56) = 1.30, p = .523$. The finding that time pressure did not lead to a shift to lower level strategies means Hypothesis 1 is not supported.

3.4.3 Test Scores
The test scores detected concealed knowledge better than chance in both the standard and time pressure conditions (see Table 1). Detection accuracy in the standard condition was $AUC = .66, p = .034, 95\% CI = [.50 .82]$, and detection accuracy in the time pressure condition was $AUC = .80, p = < .001, 95\% CI = [.67 .93]$.

**Table 3.1** Detection accuracy of total scores per condition

<table>
<thead>
<tr>
<th>Condition</th>
<th>$AUC$</th>
<th>$p$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total scores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>.66</td>
<td>.034</td>
<td>[.50 .82]</td>
</tr>
<tr>
<td>Time pressure</td>
<td>.80</td>
<td>&lt; .001</td>
<td>[.67 .93]</td>
</tr>
</tbody>
</table>

Notes. Lower scores indicate concealed knowledge.

To assess the effects of time pressure on the test scores we compared examinees with concealed knowledge between conditions per strategy levels. First we categorized the test scores into under-, over-, and chance performance. Table 2 displays the proportions for each strategy level. For level 0 and 1 strategies, the distributions of test scores were similar. For the examinees using level 2 strategies, in the standard condition 89% fell within chance performance with only 11% showing underperformance. In the time pressure condition only 37.5% fell within chance performance with 50% displaying below chance level performance. Time pressure seemed to affect only one strategy level, so we tested whether scores outside chance performance (under- and overperformance combined) were more likely to occur under time pressure for level 2 strategies. A chi-square test revealed a significant effect, $X^2 (1) = 4.90, p = 0.27$, test scores outside chance performance occurred more frequently under time pressure. Additionally, we conducted an
independent samples t-test on the absolute test scores, because not all assumptions of
the Chi squared test were met. Examinees using level 2 strategies had higher test
scores in the time pressure condition ($M = 1.90, SD = 0.62$), than in the standard
condition ($M = 0.77, SD = 0.69$), $t(15) = -3.50, p = .003$. Altogether, this supports
our second hypothesis that test scores become more extreme under time pressure
although only for examinees using level 2 strategies.

**Table 3.2** Percentage of total scores per strategy level categorized into Under-,
Chance-, and Overperformance for examinees with concealed knowledge.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Underperformance</th>
<th>Chance</th>
<th>Overperformance</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 0</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>20%</td>
<td>0%</td>
<td>80%</td>
<td>5</td>
</tr>
<tr>
<td>Time Pressure</td>
<td>33%</td>
<td>0%</td>
<td>66%</td>
<td>3</td>
</tr>
<tr>
<td><strong>Level 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>13</td>
</tr>
<tr>
<td>Time Pressure</td>
<td>78%</td>
<td>11%</td>
<td>11%</td>
<td>18</td>
</tr>
<tr>
<td><strong>Level 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>11%</td>
<td>89%</td>
<td>0%</td>
<td>9</td>
</tr>
<tr>
<td>Time Pressure</td>
<td>50%</td>
<td>37.5%</td>
<td>12.5%</td>
<td>8</td>
</tr>
</tbody>
</table>

Notes. Scores were categorized as follows: Underperformance: $x \leq -1.96$; Overperformance: $x \geq 1.96$; Chance performance: $-1.96 < x < 1.96$. Examinees that could not be categorized in any of these strategy levels. Three were excluded in the Standard condition and one in the Time pressure condition.

**3.5 Discussion**
We subjected examinees with and without concealed knowledge to a standard FCT or a modified FCT that forced examinees to respond within two seconds for each question. We introduced time pressure to the FCT paradigm to elicit cognitive load, which we expected would lead examinees with concealed knowledge to be more likely to select lower level strategies, and hence increase the detection accuracy of the FCT.

Time pressure did not affect strategy selection. The frequencies of the strategy levels between our standard and time pressure conditions did not differ and matched those found in other experiments (see Orthey et al., 2017; Orthey et al., 2018). Yet, time pressure did lead to significantly more extreme test scores in examinees with concealed knowledge who used level 2 strategies. When categorized into under-, chance-, and over-performance, more than half of those examinees fell outside the chance performance category and only a minority managed to achieve chance performance (around 37%). This stands in sharp contrast with findings in our control condition and previous research (see Orthey et al., 2017; Orthey et al., 2018), where most examinees who reported to have randomized their answers achieved test scores that fall within chance performance. Thus, even though time pressure did not affect the strategy examinees with concealed knowledge reported to have used, it did affect their ability to successfully execute these strategies.

In terms of overall detection accuracy, both the traditional FCT and the time pressure FCT detected concealed knowledge better than chance. The standard condition had a detection accuracy close to 0.70, which is within the range of previous research (Meijer et al., 2007; Orthey et al., 2017; Orthey et al., 2018). By comparison, the time pressure condition featured one of the best detection accuracies found so far, around 0.80. A likely reason for this is the reduced success-rate of level
2 strategies, resulting in more extreme test scores that are detected by the underperformance criterion. This implies that detection accuracy could additionally be increased by making effective counterstrategies harder to perform successfully.

From a theoretical point of view, these findings suggest that strategy selection is not the only component affecting the test score. In addition, examinees' ability to successfully execute their intended strategy plays a role. In this case, time pressure led examinees following a level 2 strategy to produce more extreme test scores than those not under time pressure. Other, lower level strategies were not affected, likely because they are easier to execute (i.e. either selecting only correct or only incorrect answers). That means the influence of cognitive load must be differentiated between affecting the strategy selection or strategy execution. Further disambiguation between strategy selection, the intended test outcome, and strategy execution, the actual test outcome, is needed, especially in light of various implementations of cognitive load.

Future research on the strategy selection could focus more on making it harder to discern the test’s detection mechanism through misdirection (see Kuhn, Caffaratti, Teszka, & Rensink, 2014). For example, by adding a fake polygraph to the set up of the FCT procedure in order to make examinees believe their physiological responses were recorded during the test. As a consequence, more examinees complied with the test instructions to select the correct answer alternatives (lowest level strategy) with the polygraph setup than in the control group (see Orthey et al., 2017). In a similar manner, other, more salient forms of misdirection could be used to shape examinees’ strategy selection process.
In sum, although time pressure did not affect the strategy selection of examinees with concealed knowledge, it did affect the execution of their chosen strategy, resulting in lower success rates of level 2 strategies. As such, time pressure provides an easy to implement adjustment to the FCT that will likely increase its detection accuracy.
3.6 References


*DOI: 10.3389/fpsyg.2014.01392.*


Chapter 4:

Resistance to Coaching in Forced-Choice Testing

This chapter is based on:


10.1002/acp.3443
4.1 Abstract

The Forced Choice Test (FCT) can be used to detect malingered loss of memory for specific events in the legal arena. However, empirical evidence suggests that coaching, informing the examinee about the test’s mechanism, leads to a shift in malingerers’ response strategies towards intentional randomization between correct and incorrect answers. Here we examine the validity of the runs test, alternations between correct and incorrect answers, as a criterion for malingering. 104 participants either performed a mock crime scenario or a filler task. Then participants were coached about the FCT’s test mechanics or not and subjected to a FCT. The FCT used here was designed to differentiate response patterns that appear random with those that fall within levels of chance, by manipulating the horizontal position of the correct answer alternative on each trial. Coaching lead malingerers to adopt the intentional randomization response strategy, and malingerers were detected better than chance by the runs test. Malingers, who received no coaching, were well detected by the traditional underperformance criterion and a two-step classification process of both criteria lead to incremental validity.

4 This abstract differs from the published version.
4.2 Introduction

Forced choice testing (FCT) has been used as a test to detect malingering of sensory impairment (Pankratz, Fausti, & Peed, 1975). More recently, its use has been extended to detect cases of faked memory loss (e.g., Denney, 1996; Hiscock & Hiscock, 1989; Pankratz, 1983; Van Oorsouw, & Merckelbach, 2010) and concealed information (e.g., Giger, Merten, Merckelbach, & Oswald, 2010; Meijer, Smulders, Johnston, & Merckelbach, 2007; Orthey, Vrij, Leal, & Blank, 2017; Shaw, Vrij, Mann, Leal & Hillman, 2012), from which guilty knowledge can be inferred. In the case of concealed information detection, a typical test works as follows: A suspect is presented with a series of questions about the crime. With each question, two equally plausible answer alternatives are presented; a correct and an incorrect one. For example a question such as “What was the murder weapon” could be accompanied with two answer alternatives such as “gun” and “knife”. Suspects are instructed to select the correct answer, or guess if they don’t know. Innocent suspects – who have no knowledge of the correct answers – will have to guess on each trial, and thereby choose correct answer alternatives as predicted by chance. Guilty suspects, in contrast, know which of the two alternatives is correct. To conceal this guilty knowledge, they are inclined to purposefully select the incorrect answers, leading to underperformance, i.e., the frequency with which the correct option is chosen is below chance level. Consequently, hidden knowledge is inferred from underperformance.

Previous studies have shown that FCTs have good detection rates for innocent examinees, specificity. However, the detection rate for guilty examinees, sensitivity, is modest at best. More specifically, with a specificity ranging around 95%, sensitivity ranges from 40% to 65% (Giger et al., 2010; Jelicic, Merckelbach,
& van Bergen, 2004; Meijer et al., 2007; Merckelbach, Hauer, & Rassin, 2002; Shaw et al., 2012). These validity estimates are, however, for participants who are unfamiliar with the test’s underlying rationale. Verschuere, Meijer, and Crombez (2008) showed that sensitivity is reduced considerably when participants have been informed about this rationale (i.e. coached). These authors coached half of their participants, and then submitted both naïve and coached participants to a forced choice performance test about autobiographical details. They were able to classify 58% of the naïve liars, but none of the coached liars when using underperformance (i.e., the number of correct items selected) as the criterion. Consequently, the authors conclude that forced choice performance testing is not resistant to coaching.

The finding that coached participants beat the ‘correct total’ criterion (i.e. choosing the incorrect item more often than predicted by chance) fits with the strategy description provided by Orthey, Vrij, Leal, and Blank (2017). These authors proposed that test behaviour is governed by specific strategies, and that these strategies can be categorized into different levels in accordance with Cognitive Hierarchy Theory (CHT; Carmerer, Ho, & Chong 2004). In CHT, a strategy level indicates the degree to which it anticipates any opponent’s strategy. In terms of forced choice performance testing, the test is considered the opponent and the suspect the strategist. In particular, Orthey et al. (2017) specified three strategy levels. A guilty suspect who does not anticipate anything from the test and complies with the test instructions (‘Select the correct answer, if you don’t know, guess.’) carries out a level 0 strategy. A guilty participant who assumes the test uses a level 0 strategy (i.e., compliance with test instructions) for detection therefore includes a reaction to this assumed detection strategy and executes a level 1 strategy. The most obvious reaction is to avoid correct information, which leads to underperformance.
typically seen in a substantial proportion of guilty participants. Finally, a participant who assumes the test uses a level 1 strategy (such as detection through underperformance) will use a level 2 strategy, i.e. attempt to calibrate performance within chance level. From this follows that underperformance as a detection criterion is only suitable for detecting participants who use a level 1 strategy. Coaching participants by warning them not to underperform, should elicit higher-level strategies, such as deliberate randomization.

All three strategy levels occur naturally in naïve guilty examinees. Orthey et al. (2017) found level 2 strategies to be the most prevalent and used by around 50% of their sample. This was followed by level 1 strategies, used by around 45%. Level 0 strategies were the least prevalent and occurred rarely (around 5%). Additionally, these authors linked the prevalence of strategy levels to the detection accuracy cap of the test. The total score criterion was apt at detecting underperformance in level 1 strategies, but was not designed to detect either level 0 or level 2 strategies. This shows that the detection accuracy of the test is limited to the prevalence of detectable strategies and that detection accuracy can be increased by also detecting other strategies.

Using a level 2 strategy means that examinees will attempt to produce a random sequences of correct and incorrect answers to pass the test. Yet, the correct total criterion is not the only criterion of randomness. Another criterion is the alternation rate. For example the sequence of CORRECT CORRECT CORRECT INCORRECT INCORRECT INCORRECT contains one alternation. The sequence of CORRECT INCORRECT CORRECT INCORRECT CORRECT INCORRECT CORRECT INCORRECT contains 5 alternations. Innocent examinees alternate between correct and incorrect answers on subsequent trials at a rate of 50%. Yet it is not the case for guilty
examinees. There is strong evidence suggesting that humans cannot properly reproduce randomness. When asked to generate a random response pattern, humans were found to utilize higher alternation rates than expected from true randomness (Nickerson, 2002; Wagenaar, 1972). Multiple estimates suggest that human random responding features an alternation rate of 60% as opposed to randomness’s alternation rate of 50% (see Falk & Konold, 1997). In other words, an attempted random mixture of correct and incorrect answers can be expected to exhibit more alternations than a genuine random response pattern.

Indeed, the number of alternations between correct and incorrect has been used to detect coached participants, but with limited success. Verschuere et al. (2008) only identified 21% coached liars. Similarly, Jelicic et al., (2004) – tested the number of alternations in those participants who indicated randomization as their strategy. In their sample not a single liar was identified using this test.

A potential reason for this poor detection accuracy might lie in that – as outlined above – the difference between genuine randomness (50% alternation rate) and attempted random responding (around 60% alternation rate; see Falk & Konold, 1997), is relatively small. Such a small difference requires a large test-size (i.e., number of items or questions) to become significant, and test-sizes in Verschuere et al. (2008) and Jelicic et al. (2004) may simply have been too small to detect the difference between a deliberate and random mix of answer alternatives.

In real life, including many items in forced choice performance deception detection tests may not always be feasible. The event may, for example not have enough details the investigators can verify and are exclusively known to the
perpetrator (Podlesney, 2003). If constructing large tests is not possible, another way to enhance detection accuracy is needed.

In this experiment we attempted to increase the diagnostic accuracy of the FCT procedure without requiring additional questions. Traditionally, each question in a forced choice test is presented with two answer alternatives. The position of the correct answer alternative (e.g., left or right) is determined randomly for each trial. In the current experiment, we alternate the position of the correct answer alternative between trials. On the first trial the horizontal position of the correct answer alternative would be determined randomly, for example on the right. On every subsequent trial the correct answer alternative would be presented on the opposite side of the previous trial. This way of presenting the answer alternatives allows for two types of randomized response patterns: Guilty examinees can randomize horizontally, alternating between left and right answer alternatives (which will look like a random response pattern), or between correct and incorrect answer alternatives (which produces a total score that falls within chance performance). In our design, correct/incorrect and horizontal alternations become negatively correlated. A high number of correct/incorrect alternations is associated with a low number of horizontal alternations and vice versa (e.g., always choosing the option presented on the left results in the maximum number of correct/incorrect alternations as well as the lowest number of horizontal alternations). Our idea behind this manipulation is as follows: innocent participants – whether naïve or coached – are unaware of which of the answer alternatives is correct, and will choose to randomize horizontally. As a consequence they will show a high number of horizontal alternations, corresponding to a low number of correct/incorrect alternations. Coached guilty participants are expected to employ level 2 strategies and are faced with having to choose between
producing a sequence that looks ‘random’ (high frequency of horizontal alternations) or producing a sequence where the correct total criterion falls within chance levels. Being aware of the underlying rationale of FCT will likely result in a high number of correct/incorrect alternations. In naïve guilty examinees we expect all strategy levels to occur naturally with prevalences similar to Orthey et al. (2017), and that different criteria can detect different strategies. So the total score criterion will detect the examinees who employ level 1 strategies, while the number of runs criterion will detect examinees who employ level 2 strategies.

Specifically, in the current study we investigated two questions:

i) What is the effect of coaching on the strategies guilty and innocent participants select?

ii) Can correct/incorrect alternations that are correlated with horizontal positioning discriminate guilty from innocent participants in cases of coaching?

Our hypotheses are as follows: we expect coached guilty participants to be more likely to use higher-level strategies than naïve guilty participants (Hypothesis 1), because coaching enhances their understanding of the test mechanisms and therefore aids strategy selection. Additionally, in line with previous research, we expect the correct total criterion to distinguish naïve guilty from innocent participants, but not coached guilty from innocent participants (Hypothesis 2). In contrast we expect alternations between correct/incorrect alternatives to distinguish coached guilty from innocent participants, and thus be resistant to coaching (Hypothesis 3).
4.3 Method

4.3.1 Participants

A total of 104 students (78 female) were recruited from the first year population. Students were on average 20.32 ($SD = 5.70$) years old and received course credit as compensation. Data of one participant were excluded because he did not follow the instructions. Approval from the ethics committee was obtained.

4.3.2 Procedure

First, examinees were assigned to one of two Virtual Reality (VR) simulations in a counterbalanced fashion. Their purpose was to induce crime relevant information. Half of the examinees ($N = 52$) experienced an intelligence scenario, wherein the examinee represented an intelligence officer who had to search a terrorist’s apartment for clues about an imminent attack. The other half of the examinees ($N = 52$) experienced a real estate scenario, wherein the examinee took the role of a real estate agent who explored an apartment (different from the terrorist’s apartment). Both simulations featured an interactive 3D environment that was explored from the first person perspective. Additionally, only the intelligence scenario featured interactable objects that were marked by a salient exclamation mark. Upon interaction, a window appeared that displayed a detailed picture of that object and a short descriptive text, clarifying the pictures’ content. These objects served as the crime relevant information during the following FCT procedure. In case of the intelligence scenario the simulation terminated once all objects had been interacted with, or after three minutes in the real estate scenario.
After completing the scenario, examinees were informed that they were a suspect in a police investigation about a local terrorist and had to pass a lie detection procedure. The examinees who had experienced the intelligence scenario (henceforward referred to as guilty examinees), were instructed to lie and to convince the police that they had never been in the terrorist’s apartment. Examinees who had experienced the real estate scenario (henceforward referred to as innocent examinees), were informed that they never had been to the terrorist’s apartment and that they were falsely accused. They were told that it was their task to convince the investigators that they had no knowledge of the terrorist apartment. Then examinees were randomly divided into a coached (N = 52) and naïve condition (N = 52), evenly split over the two VR scenarios. Coached examinees were provided with an advice from their attorney warning them about the mechanisms of the lie detection test (naïve examinees received no such information and directly moved on to the next part). Coached examinees received the following information:

“I know the lie detection test you will be forced to take. They will present you with questions about a crime that only the perpetrator knows the correct answer to. You will be asked to pick an answer alternative and they will instruct you to guess. They expect liars to deliberately pick the incorrect answers, to appear innocent. However, this is exactly how they identify liars. Innocent suspects are expected to actually score within levels of chance on the test.”

Subsequently all examinees were subjected to exactly the same binary FCT. First, they were informed that they would receive a number of questions and two answer alternatives per question. (One answer alternative was always correct and encountered by guilty examinees in the intelligence scenario; the other was always incorrect and unfamiliar to both guilty and innocent examinees). examinees were
forced to select one of the two answer alternatives for each question by clicking on them with the mouse and examinees were unaware of the total number of questions that would be asked. Answer alternatives were presented pictorially and their horizontal alignment (correct answer presented on the left/right side of the screen) was determined in the following way: On the first trial of the forced choice test the horizontal position of the correct answer was determined randomly. On the consecutive trials the correct answer would always be placed on the opposite side of the previous trial. This pattern was maintained for the entire test.

After completing the FCT all examinees were informed that the lie detection test was over and that they should answer the post-test questions honestly. First, they received two open questions, ‘*What did you do to appear innocent during the lie detection test?*’ and ‘*What strategy did you have in mind to make the investigator believe that you were uninvolved with the terrorist?*’. Then guilty examinees received the questions and answer alternatives again and had to indicate the correct answer for each question, which referred to the actual stimulus encountered in the intelligence scenario. This served as a memory check. Guilty examinees remembered on average 95% of the correct answers ($SD = 5.6$; worst performance = 80%).

### 4.3.3 The Forced Choice Test

The FCT featured 20 different questions about the apartment encountered in the intelligence scenario. All answer alternatives were presented pictorially. The incorrect answer in each pair was taken from a third simulation and was therefore unbeknownst to every participant. A critical assumption of these pairs was that each option was equally plausible (Doob & Kirschenbaum, 1973) to prevent deviation
from chance due to obvious/obscure answers. We used the innocent’s answers to check for biased items. Adhering to the rejection criteria used in Jelicic et al. (2004) and Merckelbach et al. (2002) all of our items were considered unbiased, because no answer alternative was chosen by more than 70% or less than 30% of the sample. Therefore, all questions were used for the analysis.

4.3.4. Design and measures

This study featured a 2 (Veracity: guilty vs innocent) x 2 (Coaching: coached vs naïve) between-subjects design with ‘correct total’ (number of correct options chosen) and ‘number of runs’ (number of alternations between correct/incorrect options plus 1) as dependent measures. Both criteria were subjected to a z-transformation according to Siegel’s (1956) formula for binomial distributions. For the correct total criterion, z scores of 0 indicate chance performance, negative z scores indicate avoidance of correct information and positive z scores endorsement of correct information. For the number of runs the same applies in terms of number of alternations between correct and incorrect answer alternatives.

Detection accuracy was measured in terms of sensitivity and specificity. Sensitivity indicates the proportion of guilty participants correctly classified and specificity indicates the proportion of innocent participants correctly classified. Sensitivity and specificity are based on a specific cut off point. For the correct total the cut off was based on the theoretical binary distribution as we expect innocent participants to inadvertently follow it. Sensitivity and specificity were computed for the conventionally used unidirectional 5% specificity cut off, as well as for 10% and
20% cut offs (e.g. Binder, Larrabee, & Millis, 2014; Van Impelen, Jelicic, Otgaar, & Merckelbach, 2017).

Cut offs for the runs criterion were computed with sample parameters of innocent participants for both conditions. There were two reasons for this choice. First, guilty and innocent examinees were expected to deviate from the binary distribution due to our manipulation, which means a cut off based on the binary distribution would not appropriately reflect the differences between guilty and innocent examinees. Second, simulating innocent population parameters was impossible due to lack of population estimates. Consequently, we acknowledge that cut off specific detection accuracy for the runs criterion may be inflated as cut offs were derived from sample parameters as opposed to population parameters. We assessed sensitivity and specificity at the unidirectional 5%, 10%, and 20% cut offs. We choose for multiple cut offs for this criterion, because it measures a different psychological process (i.e. randomization) and therefore no optimal cut off is known yet.

Additionally, we computed the incremental validity of the runs criterion in a two-step classification procedure as in Meijer et al. (2007). First the sample was subjected to the correct total criterion to detect cases of underperformance using the traditional 5% cut off. Any examinees that passed the correct total criterion were then subjected to the runs criterion, with higher alternation rates than predicted by chance being indicative of deception. Accuracy was expressed as the combined sensitivity and combined specificity.

Assessing the accuracy of such a two-step procedure is relevant, because level 2 strategies occur naturally in naïve guilty. In fact, in Orthey et al. (2017) it was
the most prevalent strategy, meaning that the runs-criterion could be relevant even for cases without coaching. Furthermore, as seen in Orthey et al. (2017) some examinees who employed level 2 strategies still were detected using the total score criterion, likely because they incorrectly judged how many correct items were required for the test score to still fall within chance performance. Therefore, we must estimate how many cases of level 2 strategies still get detected by the total score criterion, as these cases would have been detected anyway. The remaining detection accuracy then indicates the incremental validity of detecting intentional randomization. As sensitivity and specificity correspond to a specific cut off point they do not generalize to other cut offs. Instead, the Area Under the Curve (AUC) can be used as an indicator for detection accuracy independent of cut off points. It is based on the Receiver Operator Characteristic (Tanner & Swets, 1954; ROC), which plots sensitivity against specificity for the entire range of the continuous criterion. The AUC is the area covered by the ROC. It ranges between 0 to 1 with 0.5 indicating chance performance, and a higher number meaning better discrimination between guilty and innocent examinees.

Participants answers to the open questions about their behaviour during the test were categorised into three strategy levels. Level 0 strategies represented compliance with the test instructions to select the correct answers alternatives. Participants who indicated that they selected answers they thought were correct or those who indicated to use no strategy were assigned to this level. Level 1 strategies represented a reaction to the test instructions. Participants who said they avoided correct answers on purpose or controlled their demeanour while selecting answers were assigned to this level. Level 2 represented patterns that purposefully included correct and incorrect answers. Participants who said they imitated responses patterns
they believe people ignorant of the crime information would produce, or said they selected answers that seem obvious (either correct or incorrect), or indicated purposefully randomising between correct and incorrect answers were assigned to this level. Two blind and independent raters categorised the responses according to examples within each strategy level as specified in Orthey et al. (2017). Inter rater reliability was high (89% absolute agreement). Responses that did not fit any category were omitted from the analysis (1 participant).

It is important to note that the strategy level measure indicates the intended behavior of the participant only. For guilty participants the strategy level is predictive of the total score (level 0 => overperformance, level 1 => underperformance, level 2 => chance performance). For innocent participants this is not the case, as by definition they were unaware of the correct answer alternatives and the alternatives were equally plausible. As their beliefs over which particular item was correct was unrelated to the true veracity of the test items, their strategy level should be unrelated to the total score criterion. Consequently, we can assume that manipulating examinees beliefs will only have behavioural consequences for guilty examinees.

4.4. Results

4.4.1 Strategies

First we examined the strategies examinees reported. We hypothesized that coaching would elicit higher level strategies in guilty examinees (Hypothesis 1). Table 1 depicts the frequencies of selected strategies divided by conditions. Innocent examinees reported using all types of strategies naturally, but when coached they seemed to endorse either answering honestly or randomising. Naïve guilty
examinees also reported using all three strategy levels. Level 2 strategies were the most frequent, followed closely by level 1 strategies. Level 0 strategies occurred rarely. When coached guilty examinees exclusively used level 2 strategies.

Table 4.1 Frequencies of strategy levels per condition

<table>
<thead>
<tr>
<th></th>
<th>Truth tellers</th>
<th>Liars</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naïve</td>
<td>Coached</td>
</tr>
<tr>
<td>Level 0</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>Level 1</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Level 2</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>N</td>
<td>26</td>
<td>26</td>
</tr>
</tbody>
</table>

A chi-square test was performed and we found a relationship between coaching and the used strategy level for guilty examinees, $X^2 (2, N = 51) = 16.32, p < .001$. Coached guilty examinees were more likely to exhibit a level 2 strategy than naïve guilty examinees. A closer look at the data revealed that the entire sample of coached guilty examinees used a level 2 strategy, whereas the naïve guilty examinee sample consisted out a number of level 0, 1, and 2 strategies ($M = 1.44, SD = 0.65$). This supports Hypothesis 1.

Additionally, we analyzed the detection accuracy of the correct total criterion per strategy level. Ninety percent of naïve guilty examinees, who used level 1 strategies were correctly identified, whereas 23.1% of naïve guilty examinees, who
used level 2 strategies were correctly classified. All coached guilty examinees reported using level 2 strategies and only 8% of them were correctly classified. Together this supports the idea that the correct total criterion is apt at detecting level 1, but not level 2 strategies and that coaching facilitates the use of level 2 strategies.

4.4.2 Detection Accuracy

We assessed detection accuracy for specific cut-offs as well as the entire range of possible criteria (see Table 2). First we examined the correct total criterion. In the naïve condition a low correct total differentiated guilty from innocent examinees better than chance\(^5\), \(AUC = .69, p = .020, CI = [.53 .86]\). In the coaching condition the correct total did not distinguish guilty from innocent examinees better than chance, \(AUC = .53, p = .742, CI = [.37 .69]\). Similarly, when using the conventionally used unidirectional decision cut off of 5%, we found a 48% sensitivity and a 92% specificity in naïve guilty examinees. Using a 10% cut off sensitivity rose to 56% while specificity remained the same at 92.3%. At the 20% cut off sensitivity was 64% with a specificity of 88.5%. When coached, the sensitivity dropped to 7.7% with a 100% specificity at the 5% cut off. At the 10% cut off sensitivity remained at 7.7%, but specificity declined to 92.3%. At the 20% cut off sensitivity was 11.5% with a specificity of 88.5%. This suggested a sharp decline in detection accuracy for the correct total criterion in case of coaching, which supports Hypothesis 2.

\(^5\) Caution is warranted when interpreting these AUCs. The empirical ROCs are skewed (see Fig 1.), which is a consequence of the abnormal distribution of the criterion (due to different strategies used). The ROC implies that the correct total criterion is apt at detecting underperformance (level 1 strategy), but not other strategy levels. Similarly, the runs criterion performed worse than chance, because it detects over- not underperformance.
Next we examined the runs criterion. In the naïve condition, a high number of alternations resulted in worse general detection accuracy than chance\(^1\), AUC = .26, \(p = .008\), CI = [.14 .43]. However, in the coaching condition the number of runs differentiated guilty from innocent examinees significantly better than chance performance, AUC = .69, \(p = .018\), CI = [.55 .84]. We examined the detection accuracy for multiple suggested single cut offs and used the unidirectional cut offs of 5%, 10%, and 20%. In the naïve condition, the runs criterion featured a 0% sensitivity at the 5% cut off, which rose to 8% for the 10% and 20% cut off. Specificity was highest for the 5% and 10% cut offs with 92.31%. At the 20% cut off it declined to 80.71%. In the coaching condition, the 5% cut off featured a 7.69% sensitivity and 100% specificity. At the 10% cut off sensitivity increased to 34.62%, but specificity declined to 96.15%. At the 20% cut off sensitivity was 57.69% and specificity was at 69.23%. Thus, for both conditions the best sensitivity/specificity ratio was found at the 10% cut off. In any case the AUCs indicate that number of runs criterion was able to detect coached guilty examinees, supporting Hypothesis 3.
<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
<th>p</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>10%</td>
<td>20%</td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td>Total test score criterion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve</td>
<td>48%</td>
<td>56%</td>
<td>64%</td>
<td>92.3%</td>
<td>92.3%</td>
</tr>
<tr>
<td>Coached</td>
<td>7.7%</td>
<td>7.7%</td>
<td>11.5%</td>
<td>100%</td>
<td>92.3%</td>
</tr>
<tr>
<td>Number of runs criterion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve</td>
<td>0%</td>
<td>8%</td>
<td>8%</td>
<td>92.31%</td>
<td>92.31%</td>
</tr>
<tr>
<td>Coached</td>
<td>7.69%</td>
<td>34.62%</td>
<td>57.69%</td>
<td>100%</td>
<td>96.15%</td>
</tr>
</tbody>
</table>

Notes. Sensitivity & specificity for number of runs criterion were based on the undirectional 5%, 10%, and 20% cut off points corresponding to the innocent samples.
Figure 4.1 Receiver operating characteristic curve for correct total and alternation criteria for naïve and coaching conditions. Receiver operating characteristic curves in the naïve condition were aberrant. This is likely a consequence of the abnormal distribution of strategy levels used in this condition. In the coaching condition, all participants reported using the same strategy level.

Additionally, we expressed the difference between guilty and innocent examinees for the correct total and runs criterion in terms of their effect size Cohen’s $d$. However, this indicator was only computed for the coaching condition, as only in this condition the entire guilty sample utilized the same strategy level and was therefore assumed to be normally distributed. We found no effect for the correct total criterion (Cohen’s $d = -0.02$), as the coached guilty examinees ($M = -0.38$, $SD = 1.26$) matched the responses of coached innocent examinees ($M = -0.36$, $SD = 0.99$). The runs criterion had a medium effect (Cohen’s $d = -0.41$), as coached guilty examinees ($M = -0.05$, $SD = 1.18$) favored alternating between correct and incorrect answer alternatives, but coached innocent examinees prioritized alternations between horizontal positions ($M = -0.46$, $SD = 0.91$).
4.4.3 Incremental Validity

Finally we assessed the incremental validity of a two-step classification process. As step 1 we used the correct total criterion with the conventional unidirectional cut off at 5%. That is, all participants whose correct total score fell within underperformance were classified as guilty. As the second step the remaining sample was subjected to the runs criterion using the three unidirectional cut offs 5%, 10%, and 20%. Accuracy was expressed as the combined detection accuracy of steps 1 and 2. See table 3 for corresponding sensitivities and specificities. The best ratio of sensitivity/specificity was found at the 10% cut off. In the naïve condition, we found a sensitivity of 56% and a specificity of 84.62%. In the coaching condition, sensitivity was at 42.31% with a specificity of 96.15%. Combined detection accuracies indicated that sensitivity and specificity of steps 1 and 2 were additive, suggesting a unique contribution from each criterion.

Table 4.3 Detection accuracy of two step classification using total score criterion and the number of runs criterion.

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td>Naïve</td>
<td>48.00</td>
<td>56.00</td>
</tr>
<tr>
<td>Coached</td>
<td>15.38</td>
<td>42.31</td>
</tr>
</tbody>
</table>

Notes. Total score criterion (step 1) utilized unidirectional cut off of the binary distributions. The number of runs criterion (step 2) was based on the unidirectional 5%, 10%, and 20% cut off points corresponding to the innocent samples.
4.5 Discussion

We coached half of our guilty and innocent examinees and then submitted them to a FCT. In an attempt to detect coached examinees we assessed the number of runs (alternations between correct and incorrect answers) in a modified FCT. We manipulated the horizontal presentation of correct answer alternatives to alternate between trials to create a dependency between horizontal (pattern that looks random) and correct switches (pattern that falls within chance performance). If one increases, the other has to decrease. We measured detection accuracy for the number of correct answer alternatives chosen and the number of runs as well as the strategies examinees reported they used to defeat the test.

Regarding the strategies examinees reported, frequencies of strategy levels in our naïve condition closely matched those reported in Orthey et al. (2017). Coaching increased the reported strategy level for guilty examinees and coached guilty examinees exclusively reported using level 2 strategies. This is also reflected in the detection accuracy of the correct total criterion per strategy level. In naïve guilty examinees, the test detected level 1 strategies well, but not level 2 strategies. Similarly, detection accuracy for level 2 strategies in our coaching condition was very low.

The findings from this study support the idea that strategy selection is based on the beliefs one holds over the test mechanism and that strategies translate into actual test behavior (see Zvi, Nachson, & Elaad, 2012 and Zvi, Nachson, & Elaad, 2015 for similar findings a physiological concealed memory detection test). However, it is noteworthy that detection accuracies for level 2 strategies were not the same for both conditions. In our naïve condition - and in Orthey et al. (2017) -
between 23 – 50% of guilty who used a level 2 strategy were still detected as opposed to 8% in cases of coaching. A likely explanation is already provided by Orthey et al. (2017). They reasoned that as strategy onset is currently unknown, naïve guilty examinees could have started to use a level 2 strategy too late into the test, making them therefore still detectable. In our coaching condition, this problem has probably not occurred, as participants were coached before they even started the test, which means that they could have started with their level 2 strategy at the very first question.

Detection accuracy in our naïve condition matched that of other experiments, as did the decline in detection accuracy in our coaching condition for the correct total criterion. As expected in our naïve condition we found a moderate sensitivity (48%) and good specificity (92%), which matched the range of previous experiments using naïve examinees (Giger et al., 2010; Jelicic et al., 2004; Meijer et al., 2007; Merckelbach et al., 2002; Orthey et al., 2017; Shaw et al., 2012). In the presence of coaching sensitivity declined (8%), but specificity remained high (100%), matching the findings in Verschuere et al. (2008), reinforcing their conclusion that forced choice testing is not resistant to coaching when using correct total criterion.

The AUC of the runs criterion in the naïve condition suggests below chance accuracy levels. With a 10% cut off, this criterion featured a 8% sensitivity and a 92.31% specificity. This poor detection accuracy is likely a consequence of the underlying abnormal strategy level distribution. This criterion is geared towards detecting level 2 strategies, which made up only 40% of the naïve sample. Hence sensitivity is expected to be low. Furthermore, the poor AUC is explained by the substantial presence of level 1 strategies, because underperformance is negatively related to the number of runs. Selecting only incorrect answers, also means not
switching between correct and incorrect answers, which is what the runs criterion was intended to detect. Hence its’ detection accuracy is poor when alone applied to all strategy levels at once.

However, in contrast to Verschuere et al. (2008) and Jelicic et al. (2004), our runs criterion did differentiate between coached guilty and innocent examinees. We found a medium effect as guilty examinees provided responses with stronger tendencies to randomise between correct and incorrect answer alternatives, while innocent examinees were more inclined to randomise horizontally. This difference was best expressed at the 10% cut off point instead of the commonly used 5%.

We acknowledge that single cut off accuracies may be inflated as the cut offs were computed with a sample instead of population parameters and therefore may be over fitted. However, the value of the runs criterion was clearly present in the AUC in a group exclusively reporting level 2 strategies. Thus, alternations between correct and incorrect answer alternatives can discriminate coached guilty from innocent examinees, even with small test-sizes as long as a response pattern can either look ‘random’ or fall within chance performance, but not both.

The combined detection accuracy of the two-step classification process with the correct total criterion and alternations criterion suggests that the effects of each criterion are additive. Thus, each criterion captured a unique subgroup of our guilty samples. The correct total criterion was sensitive to participants using level 1 strategies (e.g. avoiding correct information) and the runs criterion to those using level 2 strategies (mixture of correct and incorrect answers). Consequently, the runs criterion provides incremental validity to the FCT paradigm by detecting intentional randomisation either occurring naturally or as a consequence of coaching.
The argument can be made that we coached examinees specifically regarding the correct total criterion, and that similarly coaching can be extended to incorporate the runs criterion as well. Nevertheless, our findings are still relevant for two reasons. First, as level 2 strategies also occur in naïve examinees, the runs criterion can increase the detection accuracy in naïve examinees. Secondly, trying to apply countermeasures for multiple criteria at once is difficult and likely taxing on cognitive resources, thus reducing the likelihood to succeed.

As for methodology, we wish to address the common critique in deception research of virtual reality applications and mock crimes. Both are often considered a threat to ecological validity in deception detection. We argue that this is not the case here. The test itself was presented and conducted just as in reality. The virtual reality mock crime simulation only served to induce crime-related information in guilty examinees. This is necessary to ensure that the assumption is met that guilty examinees recognize the correct answer alternatives. The psychological construct researched in forced choice testing is how examinees decide to choose on each trial, not how they came to know the correct answer alternatives in each trial.

Another potential concern is the validity of verbal self-reports as our measure for strategies. There has been considerable debate about the question how accurate self-reported measures are (Nisbett, & Wilson, 1977; Ericsson, & Simon, 1980; Schwarz, 1999). The concern is that human subjects may not be aware of the true reasons of their behavior and when asked about it can only produce a post hoc rationalization. To address this issue we specifically kept our questions focused on actual test behavior (i.e., ‘What did you do to defeat the test?’ instead of ‘What was your strategy to defeat the test?’). Therefore, the impact of measurement unreliability is kept to a minimum.
In sum, we found further support for the idea that guilty examinee’s test behavior is governed by a strategy selection process based on their beliefs over the test’s mechanism. We conclude that the correct total criterion is vulnerable to coaching, but coached guilty examinees can be detected using our modified runs test.
4.6 References


Chapter 5:

Eliciting Response Bias Within Forced Choice Tests to
Detect Random Responders

This chapter has been submitted as:

Orthey, R., Vrij, A., Meijer, E., Leal, S., & Blank, H. Eliciting Response Bias
Within Forced Choice Tests to Detect Random Responders
5.1 Abstract

The Forced Choice Test (FCT) can be used to detect malingered cognitive deficits. However, empirical evidence suggests that a large proportion of malingerers uses effective counter strategies to avoid detection, such as intentional random responding. To detect randomization, we investigated the runs test, which focusses on elevated alternation rates between correct and incorrect answers, and developed a new criterion. Our new criterion was defined as biased responding towards perceived (but not actual) trial difficulty. We asked 73 participants to malinger red/green blindness and subjected them to either a standard FCT procedure or a FCT with trials of varying opacity, creating the appearance of varying difficulty. Malingerers’ responses were compared to genuine chance performance. The runs test detected malingerers better than chance in the standard FCT. With varying opacity only the response bias criterion detected malingerers better than chance. Both criteria are viable methods to detect intentional randomization in the FCT.
5.2 Introduction

The Forced Choice Test (FCT) can be used to detect feigned memory loss for events (e.g. Pankratz, 1983; Denney, 1996; Bianchini, Mathias, & Greve, 2001). In a FCT, an examinee is presented with a number of questions about the event, and each question is presented with two answer alternatives of which only one is correct. The examinee is instructed to select the correct answer to each question or to guess in case they do not know. The idea behind this test is that if an examinee truly has no recollection of the event, the total test score will fall within chance levels. Malingers tend to purposefully select incorrect answer alternatives, and are more likely to obtain test scores lower than predicted by chance (so called underperformance). Similarly, FCTs can be used to detect sensory dysfunction, e.g., deafness (Pankratz, Fausti, & Peed, 1975). Here, the examinee is presented with a series of trials, on half of which a sound is presented. When asked whether a sound was played, malingerers are more likely to underreport the number of correct answers. Laboratory studies investigating the detection accuracy for underperformance in FCTs show that sensitivity – i.e. the correct detection of malingerers - varies between 40% and 60% (see Giger, Merten, Merckelbach, & Oswald, 2010; Jelicic, Merckelbach, & van Bergen, 2004; Meijer, Smulders, Johnston, & Merckelbach, 2007; Merckelbach, Hauer, & Rassin, 2002; Orthey, Vrij, Leal, & Blank, 2017; Orthey, Vrij, Meijer, Leal, & Blank, 2018; Shaw, Vrij, Mann, Leal, & Hillman, 2012) while specificity – the accurate classification of genuine performers - is around 95% (see Giger et al., 2010; Meijer et al., 2007; Orthey et al., 2017; Orthey et al., 2018; Shaw et al., 2012).

The sensitivity estimates outlined above corresponds to the prevalence of the specific strategies malingerers employ to avoid detection. Specifically, three
hierarchical strategy levels predict different types of test scores (see Orthey et al., 2017; Orthey et al., 2018). Each level is based on the belief the examinee holds over the test’s detection mechanism. Based on this belief each strategy level is associated with a distinct response strategy. Specifically, Level 0 is associated with compliance with the test instructions, which results in endorsement of correct answers, leading to overperformance. This strategy occurs rarely (< 5%; Orthey et al., 2017; Orthey et al., 2018; Shaw et al., 2012). Level 1 strategies are based on the belief the test is designed to detect level 0 strategies, resulting in a counter-response such as selecting the incorrect answers instead, which leads to underperformance. Approximately 40% of the participants report having used Level 1 strategies (Orthey et al., 2017; Orthey et al., 2018; Shaw et al., 2012). Level 2 strategies are based on the belief that the test is designed to detect level 1 strategies and predicts a counter-response, such as providing a mixture of correct and incorrect answers, so that test scores fall within chance performance. Level 2 strategies are most prevalent (around 45 – 50%; Orthey et al., 2017; Orthey et al., 2018; Shaw et al., 2012). The traditional FCT criterion focuses on underperformance (e.g. Bianchini et al., 2001; Van Oorsouw, & Merckelbach, 2010). Hence, it is well suited for detecting level 1 strategies, but not suitable for detecting levels 0 and 2 strategies. That means in order to increase the sensitivity of FCTs it is important to improve the detection rates for level 2 strategy users, as they make up the majority of malingerers.

Examinees employing a level 2 strategy attempt to simulate patterns of randomness. To detect this, the ‘runs test’ has been suggested. The criterion in this test is the number of alternations between correct and incorrect answers. It is based on the consistent finding that humans produce more alternations (∼60% alternation rate) than expected by chance (∼50% alternation rate) when trying to generate a
random sequence of two options (see Wagenaar, 1972; Falk, & Konold, 1997; Nickerson, 2002). In previous studies, the runs test yielded limited success, identifying only a fraction of malingers (Jelicic et al., 2004; Verschuere, Meijer, & Crombez, 2008). A likely reason for the poor diagnostic validity found in these studies is a lack of power (Orthey et al., 2018). The alternation likelihood of real chance performance (50%) and alternations generated by humans (≈ 60%) are too similar to elicit statistically significant differences in short tests. This systematic difference becomes visible only in tests containing a sufficient number of items. In the current study, we implement the runs test on a considerably longer FCT than in previous studies, hypothesizing that the runs test becomes diagnostic with larger test sizes.

Aside from the runs criterion, we also explore the possibility of introducing an additional criterion specifically designed to detect level 2 strategies. This idea draws on the principle of performance curves, which describe the natural decline of performance over test items with increasing difficulty (e.g. Gudjonsson, & Shackelton, 1986; Frederick, & Crosby, 2000; Frederick, Crosby, & Wynkoop, 2000; Frederick, & Foster, 1991). Frederick and Foster (1991) examined malingered cognitive deficits with a FCT of 100 trials in which the examinee had to identify relationships among abstract figures. The difficulty ranged from items so easy that even patients with genuine cognitive impairment could get the correct answer, to items so difficult that unimpaired examinees’ likelihood to select the correct answer equalled chance performance. Even though the length and slope of this performance decline may differ between individuals, they all share the same pattern, namely that performance gradually declines with increasing difficulty. Interestingly, this was not the case for malingerers, who performed worse than chance on easy items and
trended towards chance performance on items with increasing difficulty. Performance curves can also be introduced in a FCT by breaking it up into separate segments. Hiscock and Hiscock (1989) report the case of a patient whom they suspected of malingering. He was asked to memorize a five-digit number and to identify it among two alternatives after a short retention interval. The task was divided in three blocks of 24 trials with retention intervals of five seconds in the first block, ten seconds in the second block and 15 seconds in the last block. The task was designed to be so easy that the retention interval has no effect, evidenced by the performance of a five-year old, who showed above chance level performance for all three intervals. The patient displayed chance performance in the first block and below chance performance in the second and third block. Consequently, the authors suggest that malingerers adjust their test performance relative to the perceived difficulty of the test.

So far, the effect of performance curves has only been investigated for the underperformance criterion. Instead, we test a new criterion that produces a performance curve as a function of the perceived – but not the actual – difficulty, sensitive to randomizing between correct and incorrect answers. Take, for example, a standard FCT to detect malingered red/green blindness. On each trial, an examinee is presented with a red and green square, and asked to select the green one. Malingerers using level 2 strategies, would select red and green squares approximately equally often, resulting in a total score within chance performance. If we vary the opacity of the red and green objects over trials, the examinee must not only take into consideration how many correct and incorrect answers were selected, but also at what opacity. Hence, malingerers could differ from chance performance by displaying a preference to avoid/endorse correct answers relative to the perceived
difficulty of the trials. Perceived difficulty was used, because it can be introduced as an orthogonal factor to the malingered cognitive deficits. So, the task may look more/less difficult, but would have no effect on genuinely impaired performance. Malingerers are expected to be unable to have an accurate estimate of how an actually impaired examinee would respond, an effect other malingering tools such as the Structured Inventory of Malingered Symptomatology (see Merckelbach, & Smith, 2003; Jelicic, Hessels, & Merckelbach, 2006; Jelicic, Ceunen, Peters, & Merckelbach, 2011) make use of as well. Consequently, an examinee may, for example, think that on trials with strong opacity, the difference between the two objects is so clear that even red/green blind participants will perceive the difference, and select the correct alternative. This would result in a correlation between correct/incorrect answers and opacity, and this correlational response bias can serve as a new criterion specifically designed to detect intentional randomization.

In the current experiment, we asked examinees to malinger red green blindness and subjected them to one of two conditions: a standard FCT or a FCT where perceived difficulty varied per trial. Perceived difficulty was induced by varying the opacity of the stimuli over trials. We chose malingered red/green blindness for two reasons. First, perceived difficulty could be manipulated easily and objectively through opacity. Second, red/green blindness is by definition associated with chance performance, not just a steep decline in ability. Therefore, response for genuine red/green blindness could be generated through computer simulation. We evaluated three measures to detect examinees employing level 2 strategies, i.e., who employ intentional randomization of correct and incorrect answers. We only analyse examinees using level 2 strategies, and therefore expect that the number of correct alternatives selected will fail to distinguish malingered from genuine red/green
blindness (Hypothesis 1). Our FCT consists of 100 trials, which is the same test length often used to assess the human ability to generate randomness (see Wagenaar, 1972; Falk & Konold, 1997; Nickerson, 2002), and considerably larger than what has been employed in previous studies (see Jelicic et al., 2004; Verschuere et al., 2008). For that reason, we expect the runs test - based on the number of alternations between correct and incorrect - to detect malingerers using a level 2 strategy better than chance, with higher alternation rates indicating malingered performance (Hypothesis 2). Additionally, we expect biased responding as a function of the varying degree of opacity. We refer to this bias simply as *response bias*, and expect this to detect malingerers better than chance (Hypothesis 3).

## 5.3 Method

### 5.3.1 Participants

We tested 84 examinees from a university undergraduate population. Genuine red/green blindness was an exclusion criterion and zero examinees were excluded for this reason. Five examinees were excluded, because they disregarded the instructions, leaving 79 remaining. As this experiment examines examinees who naturally choose a level 2 strategy of intentionally randomizing correct and incorrect answers, we excluded all participants who reported using a different strategy. As a consequence we excluded six, leaving 37 in the Standard condition and 36 in the Opacity condition. Of these 73 examinees, 53 were female, 20 were male. Their mean age was 23.00 (SD = 6.61). Examinees were rewarded for their participation with 5 euros or course credit. Ethical approval was obtained.
5.3.2 Procedure

All examinees were instructed to feign red/green blindness. To do so we provided them with some information about red/green blindness. In essence, examinees received information that both red and green look like grey to someone with genuine red/green blindness. The information was made to look like it was derived from Wikipedia (“Color blindness”, n.d.). In addition, examinees were told that a number of tests would follow to establish whether their alleged red/green blindness was genuine. The warning was issued in order to facilitate the adoption of level 2 strategies and its’ effectiveness was reflected in the small number of excluded examinees.

The test started with two filler tasks such that the FCT was embedded into a credible task battery. First, we asked examinees to give a brief written statement indicating how red/green blindness has negatively affected them in their life. After that we administered three Ishihara plots that consist of a number of differently coloured circles. The hues are chosen in a way that colour blind and examinees without visual impairment see different numbers. Each plot was provided with two answer alternatives. One was the number people with red/green blindness would have seen and the other was the number unimpaired people would have seen (Ishihara, 2011). No data were recorded on both tasks.

Then, examinees were subjected to the FCT examination on a computer and randomly assigned to either the Standard or Opacity condition. In the Standard condition, examinees were informed that in the next part they would be presented with red and green squares and were instructed to always indicate the green square. Each trial had the same structure. First, in the middle of the screen the instruction to
select the green square was presented and at the bottom centre was a ‘next’ button located. Once examinees clicked the next button the instructions disappeared and two equal sized rectangles appeared at the top of the screen. The rectangles were in their entirety red (RGB = 255,0,0) or green (RGB = 0,255,0). Examinees could indicate their choice by clicking within the particular rectangle with a mouse. The horizontal location (left/right) of the green square was determined randomly on each trial. In total 100 of these trials were presented to each examinee.

In the Opacity condition we manipulated the opacity of both rectangles. Opacity refers to how see-through the rectangles were and can range from 100% - not see-through at all - to 0% - completely vanished -. In essence, with lower opacity it becomes harder to perceive the colour of both rectangles. Out of the 100 trials, 10 featured 100 % opacity and were identical to the trials in the control condition. The remaining 90 trials featured opacities from 10 % to 99 % in increments of 1%. We chose to omit trials with opacities lower than 10%, to make sure people with normal vision can still reasonably be expected to perceive the colour of the stimuli. The order of presentation was randomized over all 100 trials.

After the FCT, examinees were told that the assessment was over and that they should answer everything honestly. Then examinees were asked the following question: “What did you do during the test procedure to make the investigator believe that you are actually red-green color blind?” Their response was recorded, transcribed and coded by two independent coders (see below).

Finally we presented all examinees with 20 trails featuring 100% opacity. Their task was to honestly indicate the green rectangle on each trial. This served as a
performance check. Examinees who made one or more mistakes on this task were excluded. Zero participants were excluded for this reason.

5.3.3 Design

Three dependent variables were used. We computed the correct scores by summing the number of trials where the correct answer alternative was selected. For the ‘runs test’ we computed the number of alternations between correct/incorrect items. Both scores were transformed into z-scores according to the binomial distribution (see Siegel & Castellan, 1988, p. 43). Hence, the z-scores indicated how (un)likely the raw score was to occur through chance. In the Opacity condition, we estimated the response bias by conducting within each examinee a t-test for a point-biserial correlation with their choice (correct/incorrect) on each trial and the corresponding opacity used in that trial. As a result, we obtained for each examinee a correlation, indicating the strength and direction of the bias, and a p-value, indicating the significance of the correlation. We used the p-value as criterion for the response bias as the smaller the p-value was, the more unlikely the response bias was to occur through chance. The reason we chose the p-value over the correlation was that unlikely correlations can be positive or negative. The p-value avoids this issue, as it is the same disregarding the sign of the correlation. Furthermore, we computed a binary criterion, which classified whether the correlation could be considered statistically significant using the (one-sided) 5% cut off.

Examinees who are truly colour blind can be expected to show random performance. We therefore compared the distribution of our malingered examinees to a simulated random distribution. Thus, this experiment featured a 2 red/green
blindness (Malingered vs Genuine) x 2 Opacity (Standard vs Opacity) between-subjects design. We simulated the response patterns of the genuine red/green blindness group for the standard and opacity condition. Each response pattern was simulated to the trial level. By using random numbers we determined on each trial whether a participant would select a correct or incorrect response with a 50% probability each. Random numbers were generated using atmospheric noise (see RANDOM.org). With these random numbers we simulated choices as if an examinee was guessing on each trial. We computed the correct score, number of runs and the response bias the same way as for the malingerers. In total, 5 000 responses were generated for genuine red/green blindness in the Standard and Opacity condition each.

The validity of the three dependent variables will be assessed by their ability to discriminate malingered from genuine red/green blindness. General detection accuracy was assessed with the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) for the different criteria (see Tanner & Swets, 1954; Hanley, & McNeil, 1982). The ROC plots the sensitivity, detection rate for malingerers, against the specificity, detection rate for genuine performance, for all possible cut offs. The AUC ranges from 0 to 1, with 0.5 indicating chance performance. Values significantly higher than 0.5 suggest that the criterion has diagnostic value.

Examinees’ answers about their behaviour during the FCT were transcribed and coded into three strategy levels as suggested by Orthey et al. (2017) and Orthey et al. (2018). These strategy levels were referenced to the original test instruction (‘Select the correct answer alternatives. If you don’t know, guess.’) and were defined as follows: A Level 0 strategy forms no beliefs over the test’s classification
mechanism and leads to compliance with the test instructions (i.e. overperformance). A Level 1 strategy forms a belief based on the instructions and behaviour manifests as a reaction to it. The most common behaviour is intentional avoidance of correct information leading to underperformance. A Level 2 strategy is based on the belief that the test uses a Level 1 classification mechanism and therefore test behaviour manifests as a reaction to a Level 1 strategy. The most common behaviour is to attempt to provide a random mixture of correct and incorrect information.

5.4 Results

Table 5.1 Detection accuracies for all criteria for the Standard and Opacity condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Criterion</th>
<th>AUC</th>
<th>p</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>Correct total</td>
<td>.53</td>
<td>.527</td>
<td>[.42 .63]</td>
</tr>
<tr>
<td></td>
<td>Runs test</td>
<td>.69</td>
<td>&lt; .001</td>
<td>[.57 .81]</td>
</tr>
<tr>
<td>Opacity</td>
<td>Correct total</td>
<td>.38</td>
<td>.008</td>
<td>[.26 .49]</td>
</tr>
<tr>
<td></td>
<td>Runs test</td>
<td>.55</td>
<td>.299</td>
<td>[.44 .66]</td>
</tr>
<tr>
<td></td>
<td>Response Bias</td>
<td>.69</td>
<td>&lt; .001</td>
<td>[.60 .79]</td>
</tr>
</tbody>
</table>

Notes: The Correct total and Response Bias criteria assume lower scores to be indicative of malingering, while the Runs test assumes larger scores to be indicative of malingering.

Table 5.1 displays the detection accuracies using the correct scores, the runs test and the response bias as detection criteria, respectively. As hypothesized, the correct scores did not distinguish malingered from genuine red/green blindness in the
Standard condition, $AUC = .53, p = .527, 95\% \text{ CI } [.42 .63]$. In contrast, malingerers in the Opacity condition were detected with below chance level performance, $AUC = .38, p = .008, 95\% \text{ CI } [.26 .49]$. This supports our first hypothesis that the underperformance criterion has no predictive validity for examinees randomizing between correct and incorrect answers.

**Figure 5.1** AUCs of the Runs test in the Standard and Opacity condition as well as the Response bias in the Opacity condition for all test lengths from 12 to 100 trials.
The runs test detected malingerers in the Standard condition, $AUC = .69, p < .001$, $95\%\ CI [ .57, .81 ]$, but not in the Opacity condition, $AUC = .55, p = .299$, $95\%\ CI [ .44, .66 ]$, better than chance. Hence, there was only partial support for our second hypothesis that the runs test can detect examinees randomizing between correct and incorrect answers. To further estimate the relationship between test length and detection accuracy we computed the AUC for all test lengths by taking the first $n$ trials, with $n$ varying from 12 (as recommended in: Van Oorsouw, & Merckelbach, 2010) to 100 (see Figure 1). The trend suggested that in the Standard condition detection accuracy increases with test size and peaked at a test size between 50 to 70 trials. In the Opacity condition the detection accuracy of the runs test declined with test length continuously.

Finally, we assessed the validity of the response bias in the Opacity condition. We used the $p$-value as a continuous criterion as it indicates how (un)likely a response pattern is to occur through chance. The AUC was estimated using lower scores as indicative of malingering. We found that this criterion differentiated malingered from genuine red/green blindness better than chance, $AUC = .69, p < .001$, $95\%\ CI [ .60, .79 ]$. Next we estimated whether malingerers produced more response biases that pass the statistical significance threshold than expected by chance. A chi-square test indicated that malingerers were more likely to exhibit a significant correlation within their response pattern than expected by chance, $X^2 = (1, N = 5038) = 44.74, p < .001$. We found that 28.95\% of malingerers passed the 5\% threshold. Of the malingerers, who exhibited a statistical significant response bias, 64\% displayed a positive correlation ($mean = .43, SD = .17$) and 36\% displayed a negative correlation ($mean = -.32, SD = .22$). As expected, of the simulated genuine red/green blindness 4.94\% fell below the 5\% threshold. Furthermore, when
calculated over all possible test lengths (see Figure 1), the AUC of the response bias increased gradually with test length and peaked at 100 trials. These findings support our third hypothesis that the response bias can serve as a valid indicator of malingering.

5.5 Discussion

This study examined the diagnostic value of correct total scores, the runs test and the response bias criteria to detect malingered red/green blindness in examinees who utilize level 2 strategies, i.e., who randomize between correct and incorrect answers, in a Forced Choice Test. In the Standard condition all trials were identical, but in our Opacity condition we varied the opacity of both stimuli over all trials in order to tempt malingerers into adjusting their alternations between correct and incorrect answers according to the opacity of the trials. The purpose of this manipulation was to elicit an additional response bias that could serve as a new criterion to detect those who employ level 2 strategies.

The results in the Standard condition suggest that the runs test has diagnostic value provided the test size is large enough. This finding is encouraging for those applications of the FCT where the number of trials that are included in the test is unbound, such as cases of cognitive deficits. For alleged memory, auditory, or visual impairments, trials can be easily generated and repeated. It has less relevance, however, in situations where the trials are specific to unique pieces of information, for example in cases of autobiographical memory loss (e.g. Jelicic et al., 2004; Verschuere et al., 2008). Moreover, the figure plotting the validity for the different test lengths indicates a potential test length around 50 to 70 trials, after which the
accuracy of the runs test decreases. Future studies could help investigate whether this finding replicates, and help pinpoint the optimal test length for this criterion.

The effectiveness of the ‘runs test’ was limited to the Standard condition, and not present in the Opacity condition. Instead, the response bias proved a valid indicator of malingering. As seen in Figure 1, the detection accuracy of the response bias gradually increased, while the detection accuracy of the runs test gradually decreased. A potential reason for the ineffectiveness of the runs test could be that the response bias, in form of the varying opacities, is very salient and malingers preferred to calibrate their response pattern in regards to opacity, rather than with regard to their alternation rate between correct and incorrect answers. This finding is relevant, because it suggests that response biases can be elicited through perceived difficulty. This may make performance curve decision models much more resistant to countermeasures, as the malingerer first must determine whether the subsequent trials just appear more/less difficult or actually are more/less difficult for genuine impairment. Future research may also attempt to combine both types of response bias for even better detection accuracy.

Implementing a response bias to detect malingering features two challenges. (i) The introduced bias must be varied and measured objectively. In cases of alleged malingered sensory deficits such as visual or audio impairment, degrading/enhancing the stimuli can easily be done objectively. In case of malingered memory loss, the perceived importance of questions could be manipulated, but this would be challenging to do objectively. (ii) The test must contain a sufficient number of trials for the statistical assessment of the response bias. This can easily be done for malingered sensory deficits as here trials can be repeated as often as necessary. However, for malingered memory loss creating a large enough test length is very
difficult, because each trial is unique, and events often contain only few pieces of information (see Podlesney, 2003). On top of that the malingeringer must have remembered the piece of information. Thus, in terms of practical application, the response bias criterion seems best suited for malingered sensory deficits and less so for cases of malingered memory loss.

Using simulated data to represent genuine performance may raise the concern that this limits the ecological validity of our findings. Previous simulation of control group behaviour (see Betherlson, Mulchan, Odland, Miller, & Mittenberg, 2013) has been shown to be a poor reflection of real clinical samples (see Larrabee, 2014; Davis, 2018) in estimating false positive rates as a function of increasing the number of tests used to detect malingered performance. Larrabee (2014) argues that the performance of real clinical samples resembles a ceiling effect (the majority of the sample displays almost perfect performance), rather than a standard normal distribution with a mean of 0 and standard deviation of 1 as used for the simulation. We recognize these concerns, but argue that they do not apply in this case for two reasons. (i) In a FCT, by definition, stimulus pairs featured in the trials are indistinguishable for examinees with genuine impairment. Therefore, genuine performance follows the chance distribution for all three criteria, which means the test behaviour and not only can the test result be simulated. From this follows that characteristics of the sample can be expected to be representative of reality. (ii) Furthermore, a meta-analysis of the Concealed Information Test, a test that also relies on a known distribution, suggests that simulating data is even better, as it reduces sampling biases caused by small group sizes (see Meijer, Klein Selle, Elber, & Ben-Shakhar, 2014).
Another concern may be that the increase in detection accuracy is related to statistic fundamentals. With increased sample size the p-values of the t-tests for point bi-serial correlation become smaller automatically. While this is true, it is important to realize that this only applies to the malingering. Genuine guessing can be expected to always produce the same equal distribution of p-values, regardless of test length. In contrast, with increasing test length weaker effects within the malingering population yield smaller p-values. As a consequence, detection accuracy of the criterion increases, that is at least until all malingerers that do exhibit a response bias are detected. Therefore, the effect of test length on the response bias in examinees using level 2 strategies is not trivial.

In sum, our findings suggest that examinees employing level 2 strategies in a FCT can be detected by the runs test, provided the FCT features enough trials, or by varying perceived difficulty and testing for a systematic response bias. As level 2 strategies typically remain undetected and are the most common type of strategies, these new criteria can be used to increase the overall detection accuracy of FCTs.
5.6 References


Merckelbach, H., & Smith, G. (2003). Diagnostic accuracy of the structured inventory of malingered symptomatology (SIMS) in detecting instructed
malingering. *Archives of Clinical Neuropsychology, 18*, 145 – 152. DOI: 10.1016/S0887-6177(01)00191-3


Chapter 6:

General Discussion
6.1 General Discussion

The objective of this thesis was to develop a model of malingerers’ behaviour in the FCT and to use this model to increase the FCT’s detection accuracy. This thesis suggests that malingerers can be categorized in one of three subgroups defined by strategy levels. Malingerers following a level 0 strategy are expected to comply with the test instructions, leading to test scores better than chance performance, while malingerers following a level 1 strategy are expected to avoid correct answers, leading to test scores worse than chance performance. Finally, malingerers following a level 2 strategy are expected to randomize between correct and incorrect answers, leading to test scores within chance performance. In four empirical studies examinees were asked to simulate loss of memory for a specific event or cognitive dysfunction and subjected to a FCT. Examinees’ choices in terms of number of correct items selected, alternations between correct and incorrect answers, as well as additional response biases were recorded. Their ability to discriminate malingered from genuine impairment or loss of memory as well as the relationship between detection accuracy and self-reported strategies was examined.

In this general discussion, the model of malingerers’ test behaviour will be reflected upon in light of its theoretical origin, Cognitive Hierarchy Theory. Furthermore, the detection accuracy of the traditional underperformance criterion will be discussed as well as ways to improve it. Then experimental limitations will be highlighted, followed by a discussion on varying situations a FCT can be applied to and the most promising directions to improve detection accuracy, as well as alternative practical implications. This chapter will end with a brief comment on future directions and challenges for the FCT.
6.2 The Three Strategy Levels and Cognitive Hierarchy Theory

The strategy levels of this model were derived from Cognitive Hierarchy Theory (CHT; see Carmerer et al., 2004). CHT can be summarized as follows: Players in a game will decide their strategy on their belief of other players’ strategies. This involves hypothesizing what other players believe the other players will do. With that in mind a strategy is selected that is superior to the other players’ strategies. In theory, hypothesizing about other players’ beliefs can result in an endless loop as there is no certainty about the other players’ state of mind. Therefore, CHT further states that players will choose the best strategy they could devise given their available cognitive resources. That means the sophistication of a players’ strategy is limited by the available cognitive resources that player has available.

The model proposed in this thesis defines the strategy levels as the response patterns based on how the malingerer believes the FCT intends to detect malingered performance. This model is an extension of the traditional assumption that malingerers avoid correct information purposefully (e.g. Binder, Larrabee, & Millis, 2014; Denney, 1996). Not only does it specify which malingerers are detected by the underperformance criterion, i.e. level 1 strategies, but also which malingerers are not detected by this criterion, i.e. level 0 and level 2 strategies. The added benefit of this model is that knowing which malingerers remain undetected can guide research in improving the FCT.

Despite being based on CHT, this model deviated from CHT in a number of ways. First, strategies were not based on a symmetric relationship. In CHT, a game has public rules and a player’s strategy is based on their estimate of another player’s strategy. That means, a player’s behaviour is based on his/her belief of how other
players are going to act in the full knowledge that the other players also know the rules of the game and take into consideration what other players are going to do. The FCT contains only a single player (the malingering) and the rules are not public (i.e. the malingering is unaware of the detection mechanism of the FCT). Hence, the malingering’s strategy is a response pattern based on their beliefs about the FCT’s detection mechanism. Second, the (observable) number of strategy levels is limited in the FCT. In CHT the possible number of response strategies is limited by the available cognitive resources, which theoretically can result in more higher level strategies. In the FCT levels 0 – 2 predict over-, under-, and chance performance, which covers all possible test results. Even if a level 3 strategy exists, it would produce a test score that already falls within the category of a previous strategy level, making them indistinguishable behaviourally. Third, CHT does not differentiate between strategy selection and strategy execution. Results of Chapter 3 indicated that imposing cognitive load through time pressure did not affect the strategy selection process, as in reducing the frequency of higher order strategies selected, but lowered the success rate of the selected higher order strategies instead. As shown in Chapter 2 manipulating the beliefs of malingers through misdirection of reasoning did affect test responses. Hence, the distinction of strategy selection, what test score malingerers intend to achieve, and strategy execution, the actual test result, requires further disambiguation. It also provides different angles of manipulating malingerers into displaying behaviour distinct from genuine performance.
6.3 Detection Accuracy

For the evaluation of the FCT’s detection accuracy and the validity of the proposed model, the data from the control conditions of previous chapters was combined to reduce the influence of sampling biases of each study. Data from Chapters 2, 3, and 4 was combined, because these studies featured the standard FCT procedure and therefore best reflect the FCT as used in real life today.

The traditional underperformance criterion assumes that malingerers produce test scores worse than chance performance. Experiments of Chapters 2, 3 and 4 (using the traditional one-sided 5% cut off) indicated sensitivities around 40% and specificities around 95% for the underperformance criterion. These findings are in line with previous research (i.e. Giger, Merten, Merckelbach, & Oswald, 2010; Jelicic, Merckelbach, & van Bergen, 2004; Meijer, Smulders, Johnston, & Merckelbach, 2007; Merckelbach, Hauer, & Rassin, 2002; Shaw, Vrij, Mann, Leal, & Hillman, 2012; Verschuere, Meijer, & Crombez, 2008). Similarly, when using the Area Under the Curve (AUC) as general measure of detection accuracy, lower test scores differentiated malingerers from genuine impairment better than chance. In particular, our results (AUCs .72 - .80) fell within the range of previous experiments as well (Meijer et al., 2007; Shaw et al., 2012). Together, this supports previous accuracy estimates and the notion that the FCT features an excellent specificity, but lacks in sensitivity.

The limited effectiveness of the underperformance criterion found in the current thesis and studies is not surprising in light of the suggested strategy levels. The underperformance criterion, by definition, is only sensitive to level 1 strategies and the most prevalent strategy (level 2) is specifically geared towards evading the
underperformance criterion. This is demonstrated in Table 6.1 which displays the mean FCT $z$-scores per strategy levels. As expected, in the combined sample level 1 strategies are associated with very low scores, while level 2 strategies are centred around chance performance. Similarly, the distinction between strategy levels 1 and 2 becomes evident when taking the traditional definition of under- and chance performance into account. Figure 6.1 displays a histogram of malingerers’ $z$-scores per strategy level. The underperformance criterion has almost perfect detection accuracy for malingerers using a level 1 strategy. Only few malingerers using a level 2 strategy fall within underperformance levels with the majority remaining within chance performance. In sum, the traditional underperformance criterion is excellent at detection level 1 strategies, but has a poor detection rate for the remaining subgroups of malingerers, which make up the majority of malingerers. As a consequence, the sensitivity of the underperformance criterion can be expected to approach the prevalence of level 1 strategies.

**Table 6.1.** Malingerers’ average $z$-scores of number of correct answers selected combined and separated per strategy level.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>-1.48</td>
<td>2.42</td>
<td>114 (100%)</td>
</tr>
<tr>
<td>Level 0</td>
<td>2.35</td>
<td>2.79</td>
<td>8 (11%)</td>
</tr>
<tr>
<td>Level 1</td>
<td>-3.30</td>
<td>1.58</td>
<td>33 (40%)</td>
</tr>
<tr>
<td>Level 2</td>
<td>-0.81</td>
<td>1.40</td>
<td>73 (49%)</td>
</tr>
</tbody>
</table>

Notes. $z$-Scores indicate how much the observed test score deviates from chance performance. Scores $<0$ suggest underperformance; scores $>0$ suggest overperformance; scores $\approx 0$ suggest chance performance.
Figure 6.1 Histogram of malingerers’ z-transformed test scores per strategy level. Dashed lines indicate cut-off values used for classification. The one-sided 5% cut-off ($z < -1.65 = \text{underperformance}; -1.65 < z < 1.65 = \text{chance performance}; z > 1.65 = \text{overperformance}$).
A potential concern for the evaluation of the FCT’s validity is that the definition of chance performance is arbitrary. In case of the FCT, the traditionally handled 5% cut-off was likely derived from the commonly used 5% cut-off used in hypothesis testing. A possibility to avoid choosing an arbitrary cut-off could be to use the AUC as measure of detection accuracy. The AUC indicates the combined detection accuracy over all possible cut-offs. However, due to the non-normal distribution of malingerers, interpreting the AUC is not as simple as anticipated. To illustrate this, the combined empirical sample of malingerers and genuine impairment is displayed next to a simulated sample of the same groups with the same AUC (see Figure 6.2). The empirical malingerer sample (top) is abnormally distributed and the simulated malingerer sample (bottom) assumes both groups follow a normal distribution. On the right, the Receiver Operating Characteristic, plotting sensitivity against the specificity, is displayed with four specific cut-offs indicated (1%, 5%, 10%, & 20%). As demonstrated in the simulated sample gradually increasing/decreasing the chosen cut-off is associated with an equally gradual increase/decrease of sensitivity and specificity. So, a very conservative cut-off (e.g. 1%) yields a relatively low sensitivity, while liberal cut-offs (e.g. 20%) feature a relatively better sensitivity, which comes at the cost of specificity. This is not the case for our empirical sample. Due to the non-normal nature of the malingerer distribution, conservative cut-offs already feature relatively good detection accuracy. Here the 1% cut-off features almost the same sensitivity as the 20% cut-off in the normally distributed simulated sample. Furthermore, making the cut-off more liberal yields relatively small increases in sensitivity, which means the gain in sensitivity is disproportionally smaller to the loss in specificity than predicted by the simulated samples.
Figure 6.2 Distribution of z-scores for the empirical (top) and simulated (bottom) samples of malingered and genuine test performance. The left side displays the distribution overlap and the right side the Receiver Operating Curve (ROC) for each sample. Data for the simulated was generated using the effect size from the empirical sample assuming both groups were normally distributed. Markers for specific one-sided cut-offs were added to the ROC. ■ = 1%; ● = 5%; ▲ = 10%; ♦ = 20%
Therefore, the detection accuracy of the FCT should not be evaluated with a single cut-off or AUC alone. Instead, the shape of the AUC plays an important role as it can guide in identifying suitable cut-offs. The unusually high sensitivity at high specificities is a consequence of the level 1 strategies within the malingerer sample. That means, even though level 1 strategies only make up around 40% of malingerers, this subsample can be detected with high accuracy. For example, utilising the more conservative 1% cut-off, rather than the traditional 5%, would yield very similar sensitivity, but reduces the number of false positive judgements by approximately 80%. However, this also means that the detection accuracy of the underperformance criterion at high specificities is capped at the prevalence of this strategy and other approaches are needed to increase the detection accuracy even further.

### 6.4 How to improve Detection Accuracy?

Detection accuracy of the FCT could be improved through two pathways. One possibility is to promote the prevalence of level 1 strategies, because the underperformance criterion already has excellent detection accuracy for this strategy level. The results of Chapter 2 suggest that malingerers’ strategy selection can be shaped through misdirection. By making the FCT examination look like a polygraph session (through fake sensors and machinery) examinees were misdirected into believing that their physiological responses were used to infer deception. As a consequence, malingerers were more likely to comply with the test instructions (level 0 strategy), producing extreme scores that could be detected easily. In Chapter 3 the idea to influence the strategy selection process was followed up, by imposing cognitive load to promote the adoption of strategies the underperformance criterion
detects well. The expectation was that cognitive load limits malingerers’ ability to choose a level 2 strategy, thus forcing them to follow a level 1 strategy instead. Contrary to this expectation, cognitive load did not affect the strategy selection process. Instead, it affected the quality of the strategy execution. Although cognitive load did not succeed in promoting the prevalence of level 1 strategies through time pressure, increasing the prevalence of level 1 strategies should remain a viable pathway to increase the detection accuracy of the FCT. Future attempts could focus more on manipulating malingerers’ beliefs, similar to Chapter 2, to misdirect them to choose level 1 strategies. Beliefs could be influenced by manipulating the test instructions, questions, and trial design.

Another possibility to increase detection accuracy of the FCT is to add criteria sensitive to the remaining subgroups of malingerers. Level 2 strategies, intentional randomization of correct and incorrect answers, are the largest subgroup among malingerers. The ‘runs test’ is one criterion to detect randomization behaviour. In essence it indicates the frequency an examinee alternates between correct and incorrect answers. Due to the human inability to adequately reproduce randomness (see Nickerson, 2002; Wagenaar, 1972; Falk & Konold, 1997), malingerers are expected to produce more alternations than expected by chance. In Chapter 4, the ‘runs test’ can detected malingerers using a level 2 strategy in a modified short FCT procedure and in Chapter 5 the unmodified ‘runs test’ also was a valid criterion for malingering given sufficient test length. Similarly, in Chapter 5 another potential criterion was introduced: By manipulating the perceived (not real) difficulty of the FCT trials, it was possible to elicit and measure a response bias within the malingerers’ randomization behaviour. Thus, both the ‘runs test’ and the response bias were able to detect intentional randomization behaviour. Notably, the
difference between malingering and genuine performance is less pronounced for these criteria than for the underperformance criterion, meaning that the trade-off between sensitivity and specificity was worse.

Alternatively, criteria for level 1 and level 2 strategies could also be used in conjunction by implementing a two-step classification procedure (see “successive-hurdles approach” in Meehl & Rosen, 1955). For example, first the test score could be assessed for underperformance. If the test score does not fall within underperformance range, a follow up criterion sensitive to randomization behaviour could be applied. Deception is inferred if the response pattern fails on at least one of the two criteria. To maintain the predefined false positive rate, the cut-offs of both criteria must become more conservative, because adding a second criterion adds not only to the sensitivity, but also to the false positive rate.

6.5 Limitations

The empirical studies discussed in this thesis come with a number of limitations. First, the definition of strategy levels changed throughout the experiments and may still be subject to change in the future. For example, in Chapter 2 strategy levels were defined as reactions to the test instructions and other strategy levels. This meant that strategies that do not refer to the choosing behaviour during the FCT, such as only reporting to control facial expression, would still be categorized into one of the three levels. In following experiments, only responses specifically referring to the choosing behaviour were categorized. The reason for this change was that it predicted test scores better. While this has led to an improvement
of the model’s validity, it means that information on counter strategies other than choosing behaviour was lost.

Second, measuring malingerers’ strategy levels faces several challenges. Strategy levels were derived from an open-ended question about the malingerers’ test behaviour. The concern has been raised that self-reports do not make reliable data, which could cast doubt on the validity of our strategy level measure (see Nisbett & Wilson, 1977). This is not the case here, because self-reports were collected under conditions suited for this type of data (see Ericsson & Simon, 1980). Self-report measures focussed on the actual test behaviour and not on the examinees’ intentions. Self-reports of test behaviour were then recoded into strategy levels by blind independent coders. This was done to avoid measuring post-hoc rationalizations of examinees’ behaviour. Additionally, self-reports were collected immediately after the task, with appropriate debriefing, which eliminates interference through delay or intermediary tasks.

Third, with the exception of the coaching condition in Chapter 4, strategy selection was not manipulated. This has led to small and occasionally unequal sample sizes for strategy level specific tests. However, instructing malingerers what strategy level to follow can skew the detection accuracy estimate. In Chapter 4, the effects of coaching, gaining insight into the detection mechanism of the FCT, on malingerers were investigated. An interesting finding was that malingerers using level 2 strategies performed better when coached than when they developed the level 2 strategy on their own. A possible reason for this could be, that coached examinees had less doubt that they were using the correct countermeasure and therefore committing fully to their chosen strategy, or it could be that coached examinees started immediately with their counterstrategy, while examinees without coaching
have to use the first few trials to develop their strategy. Hence, inducing the strategy level per instruction may yield different success rates of the counterstrategies and therefore can lead to skewed accuracy estimates.

Fourth, in terms of ecological validity our findings only extend to university student populations. While the test situation - the computerized presentation of trials and instructions - as well as the premise - malingerers being aware of the crime details - resemble real life conditions all our results are based on a selective subcategory of humans. Therefore, care must be taken when applying the FCT to other subgroups. For example, other subgroups may have different base rates of strategy levels and/or different success rates for the various strategy levels. Consequently, detection accuracy estimates may be skewed when the FCT is applied to other subgroups. Another potential source of error may stem from the low stakes nature of the experimental setting and the lack of additional financial incentive. For example, in the Concealed Information Test (CIT), another concealed memory detection tool based on psychophysiological measurements, providing participants with a financial incentive to remain undetected leads to a significant increase in discriminant ability of the test (see Meijer, Klein Selle, Elber, & Ben-Shakhar, 2014; Table 2). Although the FCT is a cognitive test, the lack of additional financial rewards could have affected our accuracy estimates. For example, without sufficient motivation malingerers may have spend less effort and time on deciding on their response strategy, which could have lead to higher frequencies of lower level strategies than occur in real life applications. Consequently, the contribution of the underperformance criterion may be overestimated and more focus should be attributed to criteria sensitive to level 2 strategies.
Fifth, another concern for the FCT’s validity is malingerers’ memory for the event. For each trial malingerers are forced to guess, because they do not remember the correct answer anymore, or a question was formulated about a piece of information that was not encoded in the first place, noise is added to the measurement and the FCT’s diagnostic value is reduced. Therefore, natural memory decay over time becomes a crucial threat that was not assessed in the studies of this thesis as the FCT examination followed the encoding phase immediately. This has led to a very high memory performance throughout these studies for malingerers, which may not be a given in cases of delay between the event and FCT examination. However, Nahari and Ben-Shakhar (2011) have demonstrated that detection accuracy and memory performance in the FCT and CIT depends on the type of details used. They compared central and peripheral details either tested immediately following encoding or after one week. For central details, pieces of information crucial to the event, the test and memory performance remained high even after a delay. In contrast, the test and memory performance declined as a function of delay for peripheral details, pieces of information present, but not related to the event. Thus, although memory performance in the studies reported in this thesis was high, likely due to the lack of delay, this level of memory performance can also be obtained after long delays with tests based on central details. Therefore, accuracy estimates reported in these studies generalize best to FCTs of similar size based on central details only.

Finally, the prevalence of level 0 strategies - compliance with test instructions - may be artificially inflated due to the experimental situation. It is possible that due to the situation of partaking in an experiment at a university, some malingerers were unaware or realized too late when they had to start with the
deception. Consequently, some of the malingerers using a level 0 strategy may not have done so if the situation would have been clearer to them. That means the prevalence of level 0 strategies is likely inflated, though it remains unknown to what degree. However, given that these strategies occurred only occasionally in the experiments here, the impact of this problem is limited. Furthermore, any noise generated by this problem only makes the detection accuracy estimates more conservative, because none of the measures are sensitive or intended to measure level 0 strategies. Hence, detection accuracy of the FCT would be better than estimated.

6.6 Test Construction and Practical Application

Test construction should differentiate between cases of malingered loss of memory for a specific event and cognitive deficits. These types of malingered performance differ in the potentially test size that can be generated. In cases of malingered loss of memory the maximum test size is determined by the available pieces of information, as trials cannot be repeated. For many crimes the amount of available evidence is small (see Podlesney, 2003) and therefore only a small number of trials can be generated. Similarly, even if there is plentiful evidence available, the malingerer must also remember the correct answers. It is hard for the investigator to correctly estimate what pieces of information a malingerer would remember and including trials which force the malingerer to actually guess only add noise to the criterion. In contrast, the maximum test size in cases of cognitive deficits is unbound. That is, because an ability is tested and for that trials can be repeated infinitely. For example, in case of malingered red/green blindness (see Chapter 5) an examinee could be presented with a red and green otherwise identical rectangle and asked to identify the red one. With colour being the only difference between the objects, this
trial remains valid regardless of the number of repetitions. For malingered loss of memory a trial features a question with two answer alternatives. For example, ‘Which object was the murder weapon?’ could be paired with the picture of a gun and a knife. Repeating a trial such as this does not make sense, as both answer alternatives can be distinguished (even by a genuinely ignorant examinee) and there is no reason for the examinee to divert from their previous choice. These trials would violate the assumption of the FCT that all trials are independent from each other. Consequently, maximum test sizes differ per type of malingered performance.

The difference in maximum test size also has consequences for the choice of criteria and paths to improve detection accuracy. In particular, criteria such as the runs test (see Chapter 3 or Chapter 5) or a response bias (see Chapter 5) require larger test sizes to elicit meaningful differences. Furthermore, detection accuracy of the response bias in Chapter 5 increased gradually with test size, which suggests that even better detection accuracies could be achieved with longer tests. Hence, the detection accuracy of a FCT in case of malingered cognitive deficits can be increased by using additional criteria on top of the underperformance criterion. Due to the theoretically infinite test length even smaller effects can discriminate malingered from genuine performance. In cases of malingered loss of memory, the test length size is, typically, small, which means the best pathway to increase detection accuracy is to increase the prevalence of level 1 strategies, because they are well detected regardless of test length.
6.7 Innovation in Practical Application

The FCT can be applied in various situations to detect malingered performance. Regardless the type of malingering, cognitive deficits or memory loss, a high specificity is desirable for both clinical and criminal investigations. Instead, the FCT could be used as a screening tool for criminal cases with a large number of suspects. For example, if a crime was committed in a large corporation, investigators may lack the manpower to interview all employees. By using the FCT as a screening tool, for crime relevant knowledge, the large group of potential suspects could be reduced to a manageable size. This application has several benefits: (i) A more liberal cut-off can be selected, resulting in higher detection rates. Additionally, the choice of the cut-off becomes less arbitrary as the investigator can set the cut-off to match the available manpower. If there are 100 suspects and 25 interviews can be conducted, the acceptable false positive rate can be set to 25%; (ii) The impact of false positives is less severe, as the only consequence of failing the test is to be subjected to the follow-up procedure; (iii) Manpower could be saved even further by starting the follow up procedure with the least likely FCT scores first. That is, because suspects with concealed knowledge mimicking ignorance have a much higher likelihood to produce very unlikely scores than expected by chance; and (iv) The FCT can be administered easily and the test takes very little time, which are relevant constraints in such situations. Naturally, when applying the FCT as screening tool it is imperative not to attach meaning to the test outcome. All focus and conclusions should be derived from the follow-up test.
6.8 Challenges and Future Directions

Several aspects of Forced Choice Testing require further attention. One of the core assumptions of the FCT is that answer alternative pairs featured in the trials are equally plausible (Doob & Kirschenbaum, 1973). This is especially important for cases of malingered memory loss, because answer alternative pairs refer to events and are therefore not automatically equally plausible (see Frederick, Carter, & Powel, 1995). The standard validation procedure for a FCT to detect malingered memory loss is to present the questions and answer alternative pairs to a small group of examinees, who are completely ignorant to the event, and ask them to select the correct answers. The problem is defining exactly when an answer alternative pair should be considered biased. Experiments in this thesis and others (for example Meijer et al., 2007; Shaw et al., 2012) used a rule proposed by Merckelbach et al. (2002). According to this rule, pairs of which one of the two answer alternatives is selected by more than 70% are considered biased. However, there is no objective reason to set the cut-off at 70% and not for example at 75%. Similarly, there is no guideline that suggests how large the validation sample should be. This is problematic, because the rule does not differentiate instances with biased answer alternative pairs from instances that pass the threshold due to poor sampling. The former is the type of pairs that should be excluded, whereas the latter is a side effect of using small samples. Ideally, the validation sample should be as large as possible to provide the best estimate (see Law of Large Numbers), but that leads to new practical challenges. In sum, the validation process of FCT answer alternative lacks scientific scrutiny and further research is needed in order to improve the objective basis of the FCT.
Another concern is that the relationship between test size and detection accuracy of the underperformance has not been directly investigated. Some authors recommend a minimum of 12 trials in cases of malingered memory loss (Van Oorsouw, & Merckelbach, 2010), while other suggest a FCT should at least contain 25 trials (Denney, 1996; Frederick, Carter, & Powell, 1995). As seen in Table 1.1, experiments using only 12 trials do detect significant proportions of malingerers (27 - 42%; e.g. Meijer et al., 2007; Shaw et al., 2012), but the best sensitivity (≈ 60%; Jelicic et al., 2004; Verschuere et al., 2008) has been found in FCTs with 25 trials.

Here, in Chapter 5, we only explored the effect of test length on criteria sensitive to level 2 strategies, but not the underperformance criterion. Therefore, future research is needed to determine the minimum test size required and to map the relationship between detection accuracy and test size for the underperformance criterion.

So far, only the choices examinees make in the FCT have been evaluated as potential criteria for malingering. Interpreting the choosing pattern alone is challenging due to the issues outlined above. To increase the detection of malingered performance research could focus on additional measures that are independent of the response strategies e.g. mouse dynamics (Freeman, Ambady, Johnson, & Rule, 2008; Freeman, Dale, & Farmer, 2011; Monaro, Gamberini, & Sartori, 2017). In essence, these studies indicate that, if forced to make a binary selection using a computer mouse, a drift towards the correct/relevant answer alternative can be detected when selecting the incorrect answer instead. In the FCT malingerers could be expected to show a larger drift motion when selecting an incorrect answer than when selecting a correct one. Examinees with genuine impairment would not be expected to show a differential response. This measure could potentially be used as an auxiliary criterion for the FCT to reduce the false positive rates. Thus, much more
liberal cut-offs can be handled when applying the underperformance or other criteria, because deception would only be inferred in the presence of systematic mouse movements.

A final concern is that it has been neglected to investigate the influence a FCT exerts on subsequent aspects of criminal and clinical investigations. While examinees are not directly told the correct answers to the FCT’s questions, they are still exposed to the correct answers. A consequence could be that suspects in criminal investigations become aware of what information the investigator holds. This would be problematic for the Strategic Use of Evidence interviewing technique (SUE; Hartwig, Granhag, Stromwall, & Kronkvist, 2006; Hartwig, Granhag, & Luke, 2014), as it requires the investigator to strategically reveal the available evidence in order to expose the suspects’ lies. Vice versa, through educated guessing on FCT filler trials, that is trials with no correct answer alternative, the impression could be generated that implies that the investigators have more knowledge than they do. This could be beneficial for example for the Scharff technique (Granhag, Montecinos, & Oleszkiewicz, 2015; Oleszkiewicz, Granhag, & Montecinos, 2014), which is built on the idea to elicit new information from suspects, by tricking them into believing the information is already known. Hence, future research should not only concern itself with improving the FCT itself, but also investigating the influence it exerts on the surrounding criminal and clinical investigation.

6.9 Conclusion

The FCT can be used as a tool to detect malingered memory loss or malingered cognitive deficits. Three distinct response strategies have been identified
within the malingerer sample and linked to the traditional FCT criterion, underperformance, as well as other criteria such as the runs test. The model corresponds well to the data of the experiments featured here and it serves as an aid for research to develop new criteria or adjust the paradigm in order to increase the detection accuracy even further. Due to the non-normal distribution of the malingerer population, both, single cut-offs and the AUC, should be taken into account when choosing a definition of malingered performance. For example, it was demonstrated earlier that by reducing the traditional 5% cut-off to a more conservative 1% the loss in sensitivity is disproportionally smaller than to what would be expected under a normal distribution. Furthermore, two pathways were discussed to increase the detection accuracy. Either the prevalence of level 1 strategies, which are well detected by the traditional criterion, could be increased, or new criteria sensitive to the largest subgroup of malingerers, level 2 strategies, need to be developed and implemented. Which pathway is best suited to increase detection accuracy depends on the type of malingered deficit. Examiners should distinguish between test construction for cases of malingered cognitive deficits and cases of malingered memory loss. The former, refers to a loss of an ability, which has the consequence that, in theory, an infinite number of trials can be generated and parameters such as perceived difficulty can be objectively manipulated. Therefore, other criteria such the runs test or within subject response biases are well suited for this situation. In contrast, cases of malingered memory loss have a limited maximum test size, which can be problematic for criteria such as the runs test. Instead, detection accuracy could best be improved by promoting level 1 strategies. Many challenges remain for future research to address. Practical aspects, such as the relationship between test size and detection accuracy or developing objective or uniform rules to determine
biased answer alternative pairs must be addressed to aid test construction.

Furthermore, the search for strategy independent auxiliary criteria may prove a valuable addition to the existing criteria and future research should focus on the role and influence of the FCT as part of a clinical/criminal investigation.
6.10 References


Appendices

Appendix A Favourable Ethical Opinion (Chapter 2)

Science Faculty Ethics Committee

Protocol Title: SFEC 2014 – 093, Grand Theft Loot
Date received PI response from Provisional Opinion Letter: 15/12/14
Date Reviewed: 16/12/14

FAVOURABLE OPINION – SFEC 2014 – 093

Dear Mr Orthey,

Thank you for your submission for ethical review. Having completed their review, members of the Science Faculty Ethics Committee have reached a Favourable opinion of your proposed research.

Please notify the committee of any substantial amendments to the proposed procedures, send an annual report to the committee regarding study progress and a final study report once the study has concluded. Please send these to sci.fac@port.ac.uk.

Thank you for your submission and the Committee wishes you well with your study.

Dr Chris Markham – Chair of SFEC

CC –
   Holly Shepperd – Faculty Administrator

If you would like to offer any feedback on the Science Faculty Ethics Committee process please email sci.fac@port.ac.uk, to be forwarded to the Chair.
Appendix B Favourable Ethical Opinion (Chapter 3)

Science Faculty Ethics Committee
Science Faculty Office
University of Portsmouth
St Michael’s Building
White Swan Road
PORTSMOUTH
PO1 2DT

T: 023 9284 3379
ethics-sci@port.ac.uk

Date 09/09/15

Robin Orthey
Psychology
University of Portsmouth
Robin.Orthey@port.ac.uk

FAVOURABLE ETHICAL OPINION

Study Title: Memory Performance under Time Pressure
Reference Number: SFEC 2016-060 (Please quote this in any correspondence)

Thank you for resubmitting your application to the Science Faculty Ethics Committee (SEFC) for ethical review following the 2nd SFEC review dated 11/08/15, in accordance with current procedures¹.

I am pleased to inform you that SFEC was content to grant a favourable ethical opinion of the above research on the basis described in the submitted documents listed at Annex A.

Please note that the favourable opinion of SFEC does not grant permission or approval to undertake the research. Management permission or approval must be obtained from any host organisation, including the University of Portsmouth or supervisor, prior to the start of the study.

Wishing you every success in your research

Yours sincerely,

Dr Simon Kolstoe
Alternate Chair Science Faculty Ethics Committee

Information:
Prof Aldert Vrij
Holly Shawyer - Faculty Administrator

¹ Procedures for Ethical Review, Science Faculty Ethics Committee, University of Portsmouth, October 2012 (to be updated).
Statement of compliance

SFEC is constituted in accordance with the Governance Arrangements set out by the University of Portsmouth

After Ethical Review

If unfamiliar, please consult the advice After Ethical Review which gives detailed guidance on reporting requirements for studies with a favourable opinion, including, notifying substantial amendments, notification of serious breaches of the protocol, progress reports and notifying SFEC of the end of the study.

Feedback

You are invited to give your view of the service that you have received from the Faculty Ethics Committee. If you wish to make your views known please contact the administrator at ethics-sci@port.ac.uk
## ANNEX A  Documents reviewed

The documents ethically reviewed for this application (SFEC 2015-050)

<table>
<thead>
<tr>
<th>Document</th>
<th>Version</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>A - SFEC review 'time-out' Robin Orthey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B - SFEC Protocol 'time-out Robin Orthey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C - Peer review form 'time-out' Robin Orthey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D - SFEC 2015 - 050 Provisional Opinion letter with Conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E - SFEC review 'Terror investigation Robin Orthey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F - SFEC Protocol 'Terror investigation Robin Orthey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G - SFEC 2015 - 050 2nd Provisional Opinion letter with Conditions Robin Orthey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H - SFEC Protocol 'Memory performance under time pressure' Robin Orthey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I - SFEC review 'Memory performance under time pressure' Robin Orthey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J - IRB Robin Orthey-2015b</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix C Favourable Ethical Opinion (Chapter 4)

Science Faculty Ethics Committee

Protocol Title: SFEC 2015-025, Investigation in VR
Date received PI response from Provisional Opinion Letter: 05/05/15
Date Reviewed: 06/05/15

FAVOURABLE OPINION – SFEC 2015-025

Dear Mr Orthey,

Thank you for your submission for ethical review. Having completed their review, members of the Science Faculty Ethics Committee have reached a Favourable opinion of your proposed research.

Please notify the committee of any substantial amendments to the proposed procedures, send an annual report to the committee regarding study progress and a final study report once the study has concluded. Please send these to ethics-sci@port.ac.uk.

Thank you for your submission and the Committee wish you well with your study.

Dr Simon Kostoe—Alternate Vice Chair SFEC

CC—
Julie Shayer—Faculty Administrator

If you would like to offer any feedback on the Science Faculty Ethics Committee process please email ethics-sci@port.ac.uk, to be forwarded to the Chair.
Appendix D Favourable Ethical Opinion (Chapter 5)

Dear Board,

After examination of the research protocol entitled "Red or Green? Can you tell?", submitted by Robin Orthey, the Ethical Review Committee Psychology and Neuroscience (ERCNP) came to the conclusion that there are no objections to the execution of the research project as described in the said protocol with regard to the review framework used.

The applicant has been informed that:

1. Approval has been granted for a period of five years, with the possibility to prolong.
2. If the approval has been granted for a research line, each individual study within this line must be notified to the ERCNP using the form provided on the website. This does not include studies which are reviewed by a proposal committee (i.e. fMRI, EEG and TMS).
3. Changes to the approved research protocol must be submitted by the ERCNP.
4. The reference number should be mentioned in all correspondence with the ERCNP.
5. The reference number must be indicated on all advertising communications to recruit participants.

Yours sincerely,

Prof. Dr. G. Kok
Chair ERCNP

Mr. M. Schrijnemaekers
Secretary ERCNP

Prof. Dr. A.T. Sack
Board of FPN

Cc. Robin Orthey

ERCNP
Chair: G. Kok
Executive secretary: M. Schrijnemaekers

Appendix D Favourable Ethical Opinion (Chapter 5)
Appendix E UPR16 Form

FORM UPR16
Research Ethics Review Checklist
Please include this completed form as an appendix to your thesis (see the Postgraduate Research Student Handbook for more information)

Postgraduate Research Student (PGRS) Information

<table>
<thead>
<tr>
<th>PGRS Name:</th>
<th>Robin Orthey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Department:</td>
<td>Psychology</td>
</tr>
<tr>
<td>First Supervisor:</td>
<td>Alistair Vrij</td>
</tr>
<tr>
<td>Start Date: (or progression date for Prof Doc students)</td>
<td>October 2014</td>
</tr>
</tbody>
</table>

Study Mode and Route:

<table>
<thead>
<tr>
<th>Mode</th>
<th>MPhil</th>
<th>MD</th>
<th>PhD</th>
<th>Professional Doctorate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part-time</td>
<td>☒</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full-time</td>
<td></td>
<td></td>
<td>☒</td>
<td></td>
</tr>
</tbody>
</table>

Title of Thesis:

Concealed Avoidance Testing

Thesis Word Count:

33126 (excluding ancillary data)

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University’s Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study.

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:

(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at http://www.ukrio.org/what-we-do/code-of-practice-for-research/)

a) Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?

YES ☒ NO ☐

b) Have all contributions to knowledge been acknowledged?

YES ☒ NO ☐

c) Have you complied with all agreements relating to intellectual property, publication and authorship?

YES ☐ NO ☒

d) Has your research data been retained in a secure and accessible form and will it remain so for the required duration?

YES ☐ NO ☒

e) Does your research comply with all legal, ethical, and contractual requirements?

YES ☒ NO ☐

Candidate Statement:

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):

SFEC 2014 - 083
SFEC 2015 - 025
SFEC 2015 - 050
ERCPRN-175_01_01_2017

If you have not submitted your work for ethical review, and/or you have answered ‘No’ to one or more of questions a) to e), please explain below why this is so.

UPR16 – August 2015
<table>
<thead>
<tr>
<th>Signed (PGRS):</th>
<th>Date: 27th Nov 2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Signature]</td>
<td></td>
</tr>
</tbody>
</table>