Response strategies of instructed malingerers during forced choice testing

Response Strategies of Instructed Malingerers during Forced Choice Testing:

Robin Orthey

Response Strategies of Instructed Malingerers during Forced Choice Testing:

New Measures and Criteria to detect Concealed Knowledge and Feigned Cognitive Deficits

Dissertation

To obtain the degree of Doctor of Philosophy from
The University of Portsmouth and the degree of Doctor at Maastricht University, on the authority of Rector Magnificus, Prof.dr. Rianne M. Letschert in accordance with the decision of the Board of Deans,
to be defended in public on
Tuesday the 25th of June 2019 at 14:00 hours

by

Robin Orthey

Supervisors:

Professor dr. M. Jelicic

Professor dr. A. Vrij

Co-supervisor:

Dr E. Meijer

Assessment Committee:

Prof. dr. H.L.G.J. Merckelbach, Maastricht University, the Netherlands (chair)

Dr. R. Moore, University of Portsmouth, UK

Prof. dr. B. A. Schmand, Universiteit van Amsterdam, the Netherlands

Dr. B.J. Verschuere, Universiteit van Amsterdam, the Netherlands

Table of Contents

Chapter 1:

General Introduction

Pankratz, Fausti, and Peed (1975) described the classical case of a 27 year old man, who had been hospitalized multiple times for various reasons. During hospitalization the patient reported symptoms such as bilateral hearing loss, left-sided weakness, left-sided numbness, intermittent speech difficulty, and memory deficits. With the exception of the bilateral hearing loss all symptoms disappeared quickly. Furthermore, the hospital records indicated that the patient was manipulative and exaggerated symptoms to his advantage and he made repeated requests for financial compensation due to his disability. Because of the absence of medical evidence for the hearing loss and the presence of external incentives the authors conducted a test to determine whether the bilateral hearing loss was genuine or feigned. The patient was placed in front of a light that first turned red and then blue for two seconds. During one of the two light ups, determined randomly, a 1000Hz tone was presented to the left ear 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 colour 36 times. 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, meaning 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, the authors concluded that the total score was too unlikely to have occurred by chance and, instead, was the product of deliberately selecting incorrect answers. This implies that the auditory impairment was malingered.

A second seminal case was described by Denney (1996; Case 1). This author was asked to assess the competency to stand trial of a 31 year old man. The defendant was charged with armed robbery of a bank and claimed to suffer from significant memory problems caused by Lupus Erythematosus and a stroke. A hospital stay provided further

observations that suggest the defendant exaggerated physical complaints. During the observation the defendant maintained the claim of total memory loss for the robbery with the exception of information learned through the legal proceedings. This raised the suspicion that the memory loss was feigned and in order to assess the defendant's memory loss a test was created using the information from the criminal investigation. Thirty-five questions about the bank robbery were created and each question came with two answer alternatives. For example: "What did the robber say to the teller?" A: "Give me the money" or B: "I want to see the manager". The defendant was able to answer six questions through knowledge gained from the court proceedings, so they were excluded from the assessment. Of the remaining 29 questions the defendant selected the correct answer seven times. This corresponds to a *p*-value smaller than .005, which means that if the test was repeated 1000 times a patient with genuine impairment would be expected to produce less than five test results with a score like this or more extreme. Consequently, the author concluded that, given the lack of medical support for the subjective claims and motivation to feign, the defendant had memory for the event in question, and was malingering his memory loss.

Malingering and the validity of psychological examination

The two cases described above illustrate two things: that sometimes patients feign physical or psychological impairment in order to gain personal benefits, and how this feigning can be detected. Patients can feign several cognitive impairments, such as hearing loss (Pankratz, Fausti, & Peed, 1975) or blindness (Grosz, & Zimmerman, 1965). Patients can also feign general memory problems (Hiscock, & Hiscock, 1989; Pankratz, 1983) or memory loss for specific (criminal) events. Especially the latter is relevant in the forensic arena. For example, besides the case of the armed bank robbery described above, Denney (1996)

described two other cases, namely the manufacture and distribution of methamphetamine, and distribution of cocaine, in which defendants feigned memory loss for their crime. Similarly, Brandt, Rubinsky, and Lassen (1985) reported the case of a 64 year old man charged with murdering his wife, who feigned memory problems. These cases exemplify that there are patients who simulate cognitive impairment or memory loss for their personal gain. Therefore, 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.

Although the true number of patients malingering cognitive dysfunctions is unknown, it is likely to be substantial. For example, around 25% of homicide cases feature claims of partial or complete memory loss and malingered symptomatology is the best explanation for this high prevalence of amnesia claims (Cima, Nijman, Merckelbach, Kremer, & Hollnack, 2004). Current estimates of malingering vary, 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). This suggests that malingered symptomatology in forensic assessments is more than an anecdotal occurrence.

Not only is malingered symptomatology hard to detect (Rosen & Phillips, 2004) and prevalent in forensic assessments (Mittenberg et al., 2002), it has also severe consequences to society. First, Chafnetz and Underhill (2013) estimated the monetary costs due to fraudulent disability claims in 2011 at $20.02 Billion in the US alone. Second, malingered symptomatology calls into question the validity of forensic assessment. If genuine and malingered symptomatology is not differentiated, forensic assessment either becomes a useful legal defence for those unimpaired, or forensic assessment loses its credibility and genuine impairment may not receive the required legal attention. Consequently, organizations 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.

Detecting malingering using the Forced Choice Test

One of the assessment tools to detected malingering is the procedure used in the cases described above. The general rationale behind these tests is that the examinee is presented with a discrimination task and instructed to select the correct answer on each trial. Examinees with genuine impairment are unable to select the correct answer and are forced to guess. As a result they produce total test scores that fall within levels of chance. Malingerers are aware of what option is correct, and may deliberately avoid it. Consequently, malingerers' test scores are expected to be worse than chance levels. This is known as underperformance and used as criterion to determine malingered performance.

In general, this type of test is known as the "Forced Choice Test" (FCT) or "2-alternative forced choice" test (2AFC) (see Fechner, 1889). For the purpose of detecting malingered cognitive deficits Pankratz (1983) introduced the term "Symptom Validity Test" (SVT). This term has been used in the field of concealed information detection (e.g. Meijer, Smulders, Johnston, & Merckelbach, 2007; Verschuere, Meijer, & Crombez, 2008) and older malingering studies (e.g. Giger, Merten, Merckelbach, & Oswald, 2010; Jelicic, Merckelbach, & van Bergen, 2004). However, recent developments in the field of malingering have called for a change to the term "Performance Validity Test" (PVT) for better differentiation between types of malingered symptomatology (Larrabee, 2012; Larrabee, 2015). Accordingly, PVTs are used to assess whether or not an ability is impaired.

Using the examples illustrated above this can be the ability to see or hear, or acquire and recall new information. In cases of specific (criminal) events the ability to recall information from that particular time is tested. In contrast, SVTs are used to capture the exaggeration of subjective symptoms (e.g. feelings of pain). As the experiments reported in this thesis all investigate abilities, the term FCT will be used throughout. The only exception is Chapter 2 that still used the term SVT at the time of publication.

Diagnostic accuracy of the FCT

The diagnostic accuracy of the FCT is good, but leaves room for improvement. Table 1 presents an overview of 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, et al., 2010; Jelicic, et al., 2004; Meijer, et al., 2007; Merckelbach, Hauer, & Rassin, 2002; Shaw, Vrij, Mann, Leal, & Hillman, 2012; Verschuere, et al., 2008). All studies used the underperformance criterion, test scores worse than chance performance, as indicator for malingered memory loss. In addition, all studies used the same 5% cut-off point. The cut-off point marks the line between test scores considered within or outside chance performance, meaning that test scores less likely to occur by chance than 5% were considered underperformance. Only two studies (Meijer et al., 2007; Shaw et al., 2012) provide cut-off independent accuracy estimates using the Area Under the Curve (AUC). In essence, the AUC is the sum of the detection accuracy computed for all possible cut-offs and ranges from 0 to 1, with 0.5 marking chance performance and 1 perfect performance. In the FCT the AUCs range

from .70 - .87 indicating good diagnostic accuracy (Meijer et al., 2007; Shaw et al., 2012). In sum, the FCT has a reasonable detection rate of underperformance at high specificity rates.

There are two concerns regarding the FCT's current state of knowledge. One concern is that the choice of which test scores should be considered outside chance performance remains largely unexplored. Although the traditional 5% cut-off can certainly be considered conservative, featuring a low rate of false positive classifications, it is debateable whether or not an even more conservative or liberal, higher sensitivity at the cost of elevated false positive rates, cut-off is appropriate. It seems that the choice for the 5% cut-off is likely motivated by statistical convention in the field of psychology, rather than empirical observation. Furthermore, Binder, Larrabee, and Millis (2014), and Van Impelen, Jelicic, Otgaar, & Merckelbach (2017) recently suggested to use more liberal cut-offs such as 10% or 20%. Therefore, further investigation into the optimal cut-off and the (dis)advantages of conservative and liberal cut-off points in the FCT is needed.

Besides accuracy, another concern is the lack of explanations for malingerers test behaviour. Cases of underperformance can be attributed to intentional avoidance of correct answers (Pankratz, Fausti, & Peed, 1975). However, considering the sensitivity typically found in empirical studies, this explanation accounts for only half or less of malingerers. Furthermore, some studies (Merckelbach et al., 2002; Jelicic et al., 2004; Merckelbach & Van Oorsouw, 2006) feature overperformance rates between 15 – 33%. The rationale that malingerers simply avoid correct answers cannot explain the 40 – 60% false negative rates and the presence of overperformance. Hence, the variation in test scores suggests the presence of other response strategies besides simply avoiding correct information. This is supported by evidence suggesting that malingerers who understand the rationale of the FCT are better at avoiding detection (Shaw et al., 2012) and that coaching, learning the FCTs rationale before taking the test , leads to a collapse in detection accuracy of the

underperformance criterion (see Giger et al., 2010; Verschuere et al., 2008). Finally, Jelicic et al. (2004) and Shaw et al. (2012) list malingerers' self-reported response strategies. Some of the listed strategies, such as intentionally selecting correct answers or randomising between correct and incorrect answers could account for the less than perfect accuracy rates. As of now, the literature is lacking a formal conceptualization of any response strategies besides the underperformance criterion. Better knowledge about the prevalence and of the response strategies can, in turn, be used to further develop the FCT to detect these strategies.

Aims and outline of this thesis

This thesis has two objectives. The first is to map the various response strategies that malingerers use in order to defeat the FCT. The second is to use this knowledge in order to further develop the FCT to increase its diagnostic accuracy. In the experiments reported throughout the next four chapters participants will be asked to simulate 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'.

Chapter 2 classifies the various response strategies malingerers report to use in a FCT into three distinct strategy levels. These strategy levels are based on Cognitive Hierarchy Theory (Carmerer, Ho, & Chong, 2004) and places the three possible test outcomes (underperformance, test scores worse than chance, overperformance, test scores better than

chance, and chance performance) in a hierarchy of three strategy levels. To what extent the self reported response strategies match test scores is evaluated under experimental conditions.

The findings of Chapter 2 indicate two pathways to increase detection accuracy. One way, is to increase the prevalence of the avoidance strategy, because underperformance is already well detected. This idea is tested in Chapter 3, wherein cognitive load is imposed on some malingerers through time pressure.

Another way to increase the FCT's detection accuracy is to develop additional criteria that, can detect the strategies employed by the remaining 50% of malingerers, namely those who report to randomise between correct and incorrect answers. A possible criterion for intentional randomisation is the "runs test", which is based on the likelihood of alternations between correct and incorrect answers, but previous studies suggest a poor diagnostic validity (Jelicic et al., 2004; Verschuere et al., 2008). Chapter 4 explores the idea that this poor detection accuracy is a consequence of a lack of power and that the 'runs test' has diagnostic value given a small change in the presentation style of the FCT. Furthermore, in Chapter 5 the 'runs test' and a new criterion are evaluated to detect malingered red/green blindness. By introducing perceived (not real) difficulty to the FCT's trials, malingerers are tempted to adjust their randomisation pattern to the perceived trial difficulty. Both criteria have diagnostic value in detecting randomisation behaviour.

Finally, Chapter 6 features an evaluation of the detection accuracy of the FCT, as well as a critical reflection on how to best measure its detection accuracy. In addition, the model and three strategy levels proposed in Chapter 2 are evaluated in light of all data combined. Both pathways to increase detection accuracy are then discussed in light of two distinct types of application: malingered loss of memory of a specific event and malingered cognitive

impairment (i.e. general memory loss, colour blindness). This chapter end with a reflection on experimental limitations and a possible new mode of application as a screening tool.

Table 1 Overview of the detection accuracy of the 2 alternative Forced Choice Test for malingered performance

Study	Critical (Total) FCT items	Sensitivity (N)	Specificity (N)	AUC	Malingerers' test response distribution (N)		
					Under-performance	Chance performance	Over-performance
Merckelbach et al., 2002	15 (15)	40% (20)	-	-	40% (8)	30% (6)	30% (6)
Jelicic et al., 2004	25 (50)	59% (39)	-	-	59% (23)	26% (10)	15% (6)
Van Oorsouw & Merckelbach, 2006*	21 (40)	7% (27)	100% (30)	-	7% (2)	60% (16)	33% (9)
Meijer et al., 2007 – Study 1	12 (12)	27% (30)	100% (30)	.70	27% (8)	-	-
Meijer et al., 2007 – Study 2	12 (12)	- (60)	- (60)	.87	-	-	-
Verschuere et al., 2008 – Naïve	25 (35)	58% (19)	-	-	58% (11)	-	-
Verschuere et al., 2008 – Coached	25 (35)	0% (19)	-	-	0% (0)	-	-
Giger et al., 2010 – Naïve	19 (38)	45% (20)	90% (20)**	-	45% (9)	-	-
Giger et al., 2010 – Warned	19 (38)	10% (20)	90% (20)**	-	10% (2)	-	-
Shaw et al., 2014	12 (12)	42% (86)	93% (82)	.79	42% (36)	57% (49)	1% (1)

Notes. Study * should be interpreted with care as the FCT featured many biased item pairs. Specificity marked with ** refers to the same group. Giger et al., 2010 featured three groups only: naïve malingerer, warned malingerer, and examinees without crime knowledge (marked with **).

Chapter 2:

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

This chapter is based on:

Orthey, R., Vrij, A., Leal, S., & Blank, H. (2017). Strategy and misdirection in forced choice memory performance testing in deception detection. *Applied Cognitive Psychology, 31(2),* 139-145. *DOI: 10.1002/acp.3310*

Abstract

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

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: (H1) Liars who use level 1 strategies will be detected, but liars who use level 2 strategies will not be detected; (H2) Liars will report higher order strategies than truth tellers**;** and (H3) 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 underformance. 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: (H4) 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; (H5) We expect examinees in the misdirection condition to use physiological countermeasures as their strategy to beat the test; and (H6) We expect liars in the misdirection condition to produce more cases of overperformance (significantly more questions correct than expected by chance) than liars in the control condition.

Method

Participants

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

Material

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

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

Procedure

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

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

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

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

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

After the test participants were notified that the lie detection test was over and were asked to answer the following questions honestly: 'What did you do to appear innocent on the test?' and 'Did you believe that your sweat was measured during the test? (Yes/No)'. 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.

Design

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

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

Participants' strategies were extracted from the open question 'What did you do to appear innocent on the test?' The primary investigator first read through all responses and

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

A second rater, blind to the hypothesis, classified each participant according to these eight categories. If a response would have fitted into more than one category, the coder was instructed to choose the one with the best fit. In cases of disagreement both coders discussed the instance and coded the case independent from each other again. Inter-rater reliability was good (73.9% absolute agreement).

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

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

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

Results

Understanding & Misdirection

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

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

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

Strategies

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

sophisticated than level 0 strategies (H3). Then we will address the detection accuracy of the different strategy levels. We expect the test to reliably detect level 1 strategies, but not level 2 strategies (H1).

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

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

Table 1 Self reported strategies distinguished between conditions and strategy levels.

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

Notes. Frequency of strategies indicated in percentages.

Next, we examined the strategy levels. A 2 (Veracity) x 2 (Misdirection) between-subjects ANOVA was conducted with the strategy level as dependent variable. There was a significant difference for the Veracity main effect, $F(1,82) = 10.65$, $p = .002$, $\eta^2 = .12$. Liars (*mean* = 1.50 [1.28 1.72]) used on average a higher level strategy than truth tellers (*mean* = 0.98 [0.75 1.20]), supporting H2. The main effect for Misdirection, $F(1,82) = 1.02$, $p = .315$, $\eta^2 = .01$, and the interaction between Veracity and Misdirection, $F(1,82) = 0.02$, $p = .904$, $\eta^2 < .001$, were not significant.

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

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

Table 2 Number of detected and undetected strategies differentiated by level for Liars in the Control condition (unidirectional) and Misdirection condition (bidirectional).

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

Notes. Detection accuracy indicated in percentages.

To follow up we examined in particular which level 2 strategies of liars exactly were detected and which remained undetected. Table 3 displays these frequencies for the Control and Misdirection condition together, as both showed almost the same pattern. As Table 3 shows, each of three level 2 strategies was as frequently detected as it remained undetected. In addition we looked into the individual z-scores of these participants. A considerable proportion (33.33%) of detected liars using level two strategies had just enough answers wrong to be detected.

Table 3 Frequency of liars' level 2 strategies

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

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

Avoidance behaviour & Detection accuracy

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

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

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

Table 4 Classification Accuracy

	Sensitivity	Specificity	AUC	95% CI
Control				
Unidirectional	56.52	86.95	.76*	.62 - .90
Bidirectional	60.87	86.95	.72*	.56 - .88
Misdirection				
Unidirectional	43.5	100	.67	.49 - .84
Bidirectional	65.22	95.65	.82*	.69 - .95

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

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). Fist, 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 to 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.

Chapter 3:

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

This chapter is based on:

Orthey, R., Palena, N., Vrij, A., Meijer, E., Leal, S., Blank, H., & Caso, L. (pending minor revision). Effects of Time Pressure on Strategy Selection and Strategy Execution in Forced Choice Tests. *Applied Cognitive Psychology*

Abstract

Although the Forced Choice Test (FCT) has successfully been used to detect malingered memory loss, their sensitivity leaves room for improvement. We tested the hypothesis that inducing cognitive load during FCT would impair examinees' ability to choose effective countermeasures and thereby increase the FCT's detection accuracy. We subjected 120 examinees with or without concealed knowledge about a mock crime to either a traditional FCT, or a FCT that had to be completed under cognitive load, i.e. time pressure. In this latter condition, examinees had to respond to each question within two seconds. Time pressure lowered the success rates of effective counterstrategies (but not their incidence rates). Effects of time pressure on detection accuracy and theoretical implications are discussed.

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).

Method

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.

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%).

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.

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' 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.

Results

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.

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 1 Detection accuracy of total scores per condition

Condition	*AUC*	*p*	*95% CI*
Total scores			
Standard	.66	.034	[.50 .82]
Time pressure	.80	< .001	[.67 .93]

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

Condition	Underperformance	Chance performance	Overperformance	N
Level 0				
Standard	20%	0%	80%	5
Time Pressure	33%	0%	66%	3
Level 1				
Standard	100%	0%	0%	13
Time Pressure	78%	11%	11%	18
Level 2				
Standard	11%	89%	0%	9
Time Pressure	50%	37.5%	12.5%	8

Notes. Scores were categorized as follows: Underperformance: x <= -1.96; Overperformance: x >= 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.

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.

Chapter 4:

Resistance to coaching in forced-choice testing

This chapter is based on:

Orthey, R., Vrij, A., Meijer, E., Leal, S., & Blank, H. (2018). Resistance to coaching in forced-choice testing. *Applied Cognitive Psychology,* 1 – 8. *DOI: 10.1002/acp.3443*

Abstract

In Forced Choice Tests (FCT), examinees are typically presented with questions with two equally plausible answer alternatives, of which only one is correct. The rationale underlying this test is that guilty examinees tend to avoid relevant crime information, producing a non-random response pattern. The validity of FCTs is reduced when examinees are informed about this underlying rationale, with coached guilty examinees refraining from avoiding the correct information, but trying to provide a random mix of correct and incorrect answers. To detect such intentional randomization a 'runs' test – looking at the distribution of the number of alternations between correct and incorrect answers – has been suggested, but with limited success. We designed a runs test based on distinguishing between patterns that look random andpatterns that are random. Specifically, we alternated the horizontal presentation (i.e. presentation left or right on the screen) of the correct answer alternative between each trial. As a consequence, guilty examinees were faced with having to chose to randomise either between correct and incorrect answers - leading to chance performance - or between answers presented on the left or right, producing a pattern that 'looks' random. As innocent examinees are unaware of the correct answers they can only randomise between horizontal positions. Results showed that the number of correct items selected distinguished guilty from innocent examinees only when they were not informed about the underlying rationale. In contrast, alternations between correct and incorrect answers did distinguish informed guilty from innocent examinees. Incremental validity of the alternations criterion and theoretical implications are discussed.

Introduction

Forced choice testing (FCT) has been used as a test to detect malingering of sensory impairment (Pankratz, Fausti, & Peed, 1975). More recently, it's 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 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 outline 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).

Method

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.

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%).

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.

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 74ategori during the test were 74ategorized 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 74ategorized between correct and incorrect answers were assigned to this level. Two blind and independent raters 74ategorized 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 => chanceperformance). 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.

Results

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 1 Frequencies of strategy levels per condition

	Truth tellers		Liars	
	Naïve	Coached	Naïve	Coached
Level 0	8	15	2	-
Level 1	12	1	10	-
Level 2	5	10	13	26
Other	1	-	-	-
N	26	26	25	26

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.

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[1], 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

[1] 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.

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.

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 2 Detection accuracy for the alternations criterion

	Sensitivity			Specificity			AUC	*p*	95% CI
	5%	10%	20%	5%	10%	20%			
Total test score criterion									
Naïve	48%	56%	64%	92.3%	92.3%	88.5%	.69	.020	[.53 .86]
Coached	7.7%	7.7%	11.5%	100%	92.3%	88.5%	.53	.742	[.37 .69]
Number of runs criterion									
Naïve	0%	8%	8%	92.31%	92.31%	80.71%	.26	.008	[.14 .43]
Coached	7.69%	34.62%	57.69%	100%	96.15%	69.23%	.69	.018	[.55 .84]

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.

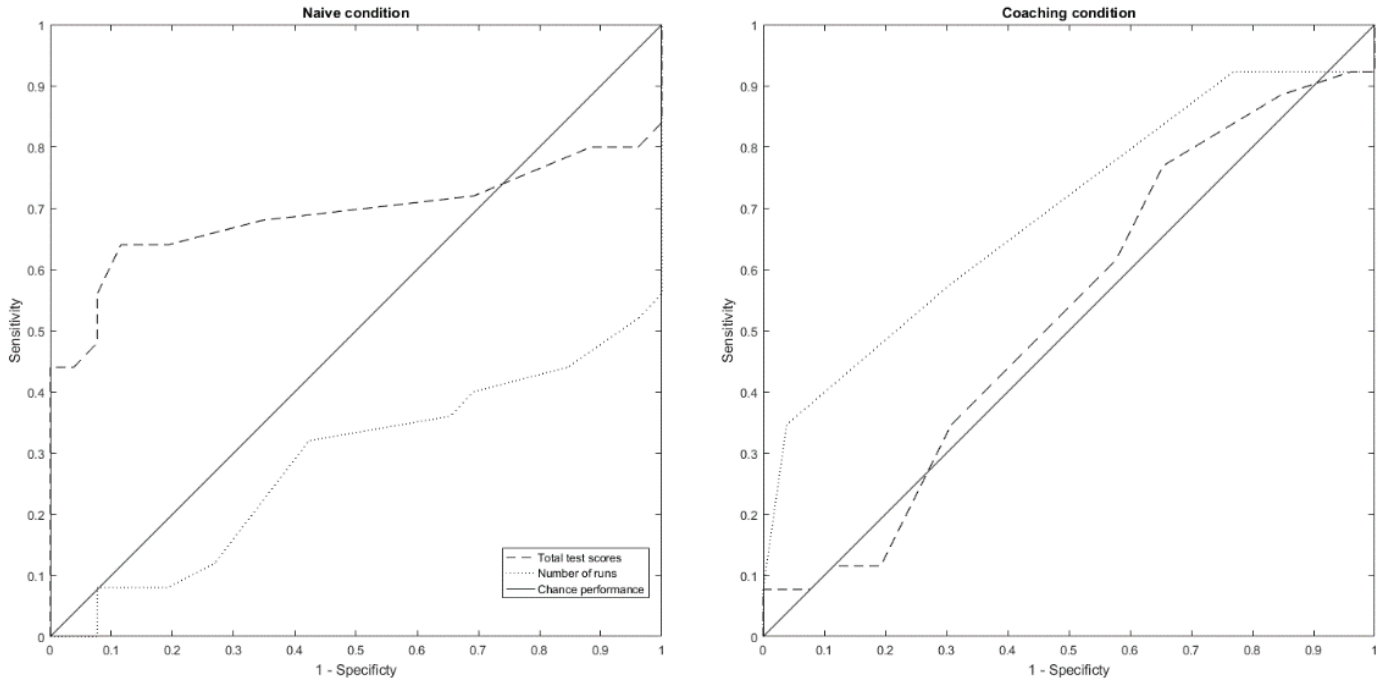

Figure 1 Receiver operating characteristic curve for correct total and alternation criteria for naïve and coaching conditions. Receiver operating characteristic curves in the naïve condition were aberrant. This is likely a consequence of the abnormal distribution of strategy levels used in this condition. 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).

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

	Sensitivity			Specificity		
	5%	10%	20%	5%	10%	20%
Naïve	48.00	56.00	56.00	84.62	84.62	73.08
Coached	15.38	42.31	65.38	100	96.15	69.23

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 undirectional 5%, 10%, and 20% cut off points corresponding to the innocent samples.

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.

Chapter 5:

Using Bias for Good. Eliciting Response Bias Within Forced Choice Tests to Detect Random Responders

This chapter is based on:

Orthey, R., Vrij, A., Meijer, E., Leal, S., & Blank, H. (2019). Eliciting Response Bias Within Forced Choice Tests to Detect Random Responders, *Nature Scientific Reports*

Abstract

The Forced choice test (FCT) can be used to detect malingered loss of memory or sensory deficits. Examinees are presented with questions with correct and incorrect answer alternatives. Genuine performance, on the other hand, is associated with test scores that fall within chance performance. A substantial proportion of malingerers intentionally randomize their responses, resulting in false negative test outcomes. Here we examine whether a 'runs test' and a 'within test response bias' have diagnostic value to detect this intentional randomization. We instructed 73 examinees to malinger red/green blindness and subjected them to a FCT. For half of the examinees we manipulated the ambiguity between answer alternatives over the test trials in order to elicit response biases. The responses of malingerers were compared to a sample of 10,000 simulated genuine performances. Results indicate that individual response bias has diagnostic value to detect intentional randomization.

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. Malingerers 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 malingerers (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).

Method

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.

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.

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.

Results

Table 1 Detection accuracies for all criteria for the Standard and Opacity condition.

Condition	Criterion	AUC	*p*	95 % CI
Standard				
	Correct total	.53	.527	[.42 .63]
	Runs test	.69	< .001	[.57 .81]
Opacity				
	Correct total	.38	.008	[.26 .49]
	Runs test	.55	.299	[.44 .66]
	Response Bias	.69	< .001	[.60 .79]

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 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% CI* [.42 .63]. In contrast, malingerers in the Opacity condition were detected with below chance level performance, *AUC* = .38, *p* = .008, *95% CI* [.26 .49]. This supports our first hypothesis that the underperformance criterion has no predictive validity for examinees randomizing between correct and incorrect answers.

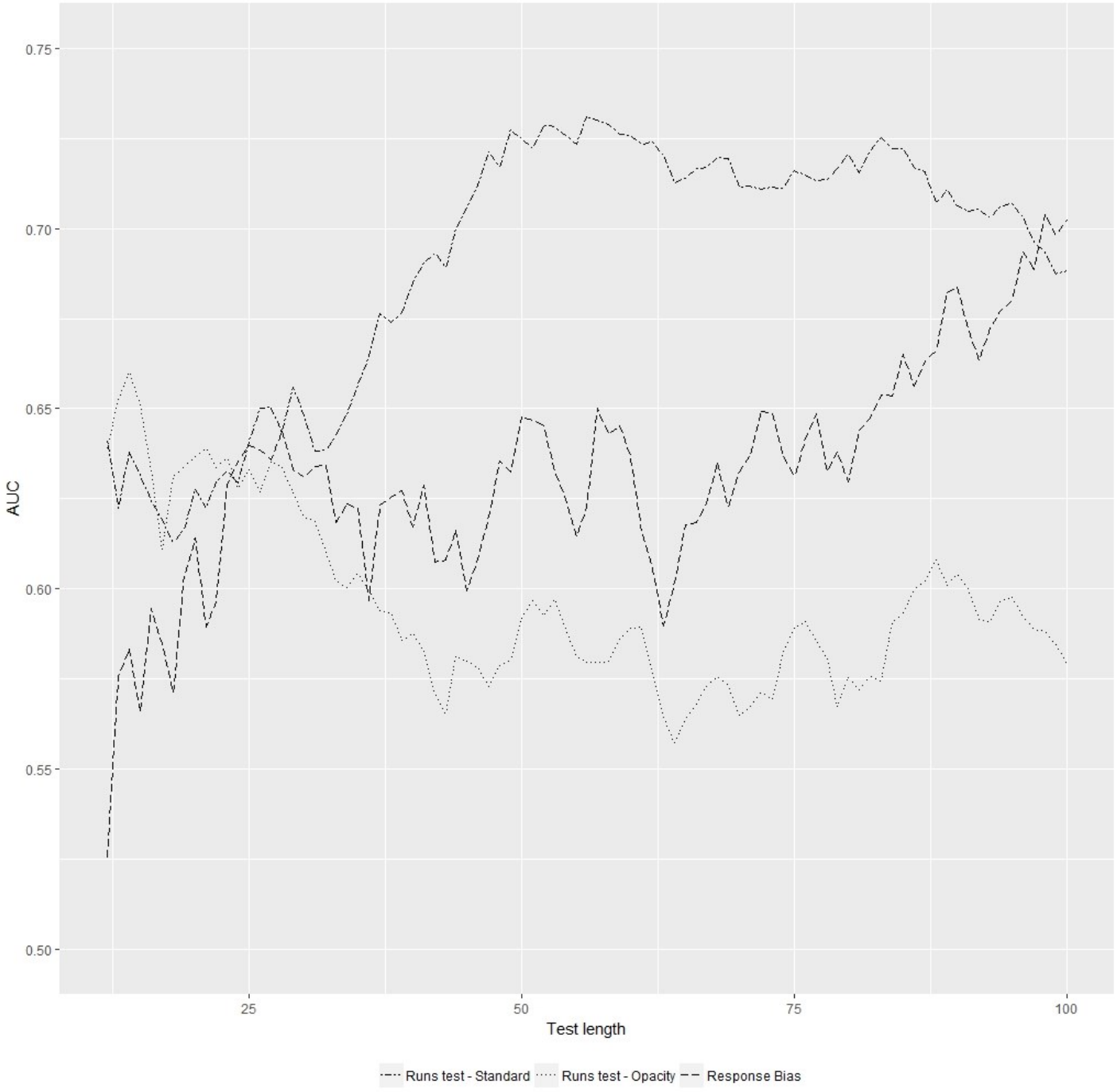

Figure 1 AUCs of the Runs test in the Standard and Opacity condition as well as the Response bias in the Opacity condition for all test lengths from 12 to 100 trials.

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.

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 malingerers 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 malingerer 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 malinger. 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 malingerer 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.

Chapter 6:

General Discussion

The objective of this thesis was to develop a better understanding of malingerers' behaviour in the FCT and to use this knowledge to increase its detection accuracy. Malingerers' countermeasures can be categorized in one of three subgroups defined by strategy levels. Malingerers following a level 0 strategy comply with the test instructions, leading to test scores better than chance performance, while malingerers following a level 1 strategy avoid correct answers, leading to test scores worse than chance performance. Finally, malingerers following a level 2 strategy randomize between correct and incorrect answers, leading to test scores within chance performance. The detection accuracy of the FCT's underperformance criterion closely matched the prevalence of level 1 strategies. The subgroup of malingerers following a level 2 strategy, was not well detected by the underperformance criterion and formed the largest source of error in classification.

In Chapter 2 malingerers' various countermeasures to the FCT were classified into three distinct strategy levels based on Cognitive Hierarchy Theory (Carmerer, Ho, & Chong, 2004). The match between self-reported strategies and malingerers' test scores was evaluated under experimental conditions. The underperformance criterion detected followers of a level 1 strategy well, but not those following a level 2 strategy.

One way to increase detection accuracy is to increase the frequency of level 1 strategies, as they are well detected by the underperformance criterion. In Chapter 3, cognitive load was induced through time pressure in order to limit malingerers' ability to devise effective counterstrategies to the FCT and promote lower level strategies. The frequency of self-reported strategy levels did not differ between malingerers put under time pressure or not. However, a notable finding was that under time pressure, malingerers, who reported to randomize between correct and incorrect answers, produced more scores considered outside chance performance. This suggests a dissociation between participants' self-report and their behaviour.

Another way to increase detection accuracy is to focus on the subgroup of malingerers that is not well detected by the underperformance criterion. In Chapter 4, the 'runs test', alternations between correct and incorrect answers, was examined in case of coached malingerers. A modified version of the runs test was created for this experiment. Here the position (left or right) of the correct answer alternative on each trial was alternated on each trial. As a consequence, genuine guessing was anticorrelated to alternations between correct and incorrect answers. The results indicate that coached malingerers could be detected by the runs test better than chance. In Chapter 5, examinees were asked to malinger red/green blindness. In addition to the runs test, a new criterion, a response bias to perceived (but not actual) trial difficulty, was tested. The results indicate that both the runs test and response bias could detect malingered performance better than chance.

To synthesize these findings, the classification of malingerers' test behaviour will be evaluated in light of its theoretical origin, Cognitive Hierarchy Theory (CHT). 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.

The three strategy levels and Cognitive Hierarchy Theory

The strategy levels of the model defined in Chapter 2 were derived from Cognitive Hierarchy Theory (CHT; see Carmerer et al., 2004). According to CHT 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 a player has available.

The strategy levels proposed in Chapter 2 define the response patterns based on how malingerers believe the FCT intends to detect malingered performance. These strategy levels are an extension of the traditional assumption that malingerers avoid correct information purposefully (e.g., Denney, 1996; Binder, Larrabee, & Millis, 2014). Not only do they 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 is that knowing which malingerers remain undetected can guide research intended to improving the FCT.

Despite being based on CHT, the strategy levels deviated from CHT in a number of ways. First, strategies are 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. This means that, 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 malingerer) and the rules are not public (i.e. the malingerer is unaware of the true detection mechanism of the FCT). Hence, the malingerer'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 behaviourally indistinguishable. 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 prevalence of self-reported strategies. However, a considerable proportion of malingerers reported to provide a mixture of correct and incorrect answers, but still produced test scores below chance performance. This discrepancy between strategy and test performance did not occur in the control condition of this experiment, suggesting that time pressure may have had an effect on examinees' ability to correctly execute their chosen strategy. Similarly, manipulating the beliefs of malingerers through misdirection of reasoning did also affect test responses (Chapter 2). 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.

Detection Accuracy

The evaluation of the FCT's detection accuracy and the validity of the strategy levels uses the data from the control conditions of previous chapters to reduce the influence of sampling biases of each study. Data from Chapters 2, 3, and 4 are combined, because the studies in those chapters featured the unmodified FCT procedure, had a similar test length, and dealt with cases of malingered loss of memory.

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 sensitivity around 54% (N = 74) and specificity around 92% (N = 74) 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 at high levels of specificity the FCT has modest sensitivity.

The limited effectiveness of the underperformance criterion found in the current thesis and previous studies is not surprising in light of the proposed 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 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 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

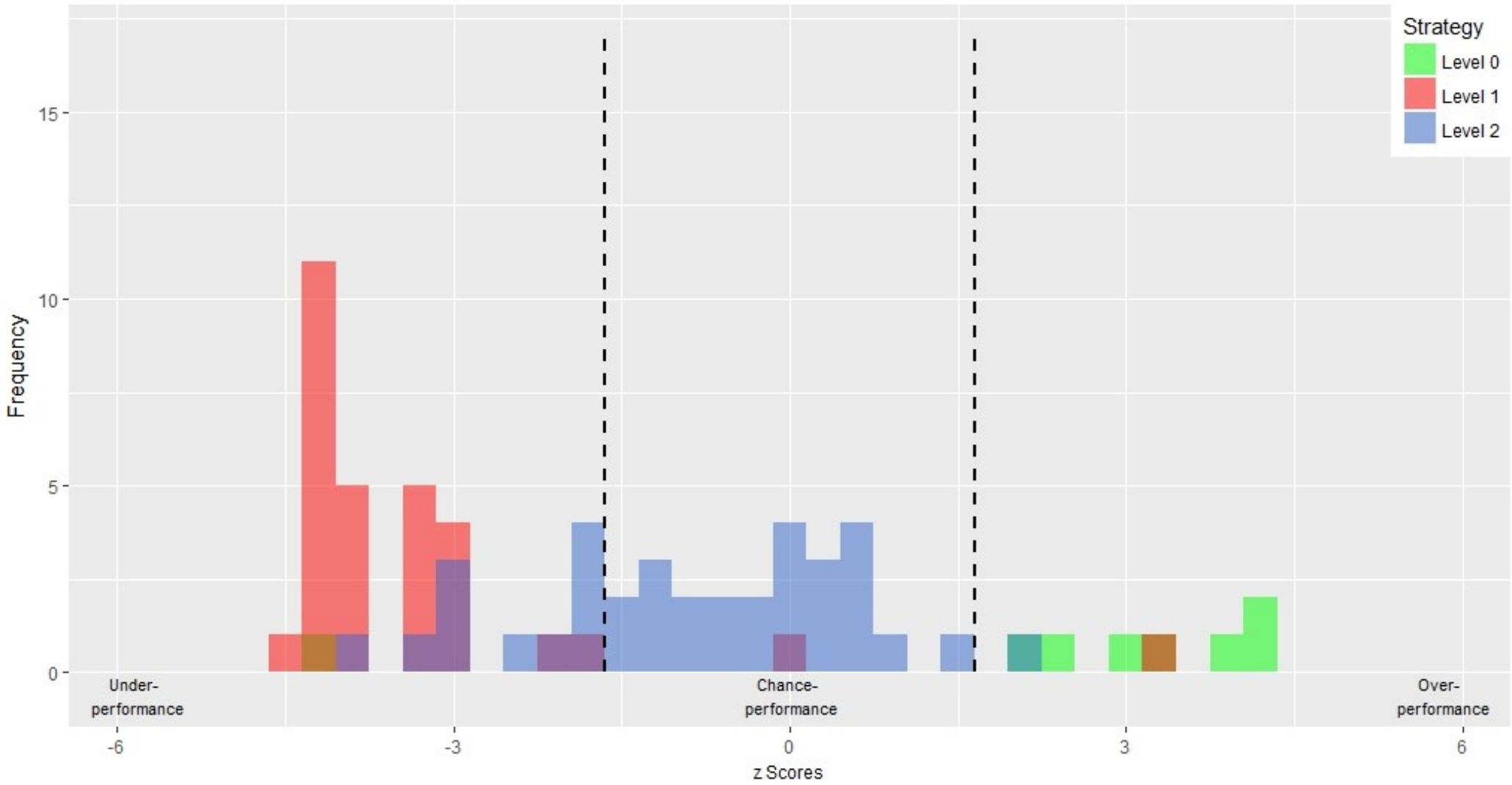

Figure 1 Histogram of malingerers' z-transformed test scores per strategy level. Dashed lines indicate cut-off values used for classification. The one-sided 5% cut-off ($z < -1.65$ = underperformance; $-1.65 < z < 1.65$ = chance performance; $z > 1.65$ = overperformance). (Unspecified colours are a result of overlap. Brown = Red and Green; Purple = Red and Blue; Turquoise = Green and Blue)

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 1 Malingerers' average z-scores of number of correct answers selected combined and separated per strategy level.

	Mean	SD	N
All	-1.48	2.42	74 (100%)
Level 0	2.35	2.79	8 (11%)
Level 1	-3.30	1.58	30 (40%)
Level 2	-0.81	1.40	36 (49%)

Notes. z-Scores indicate how much the observed test score deviates from chance performance. Scores < 0 suggest underperformance; scores > 0 suggest overperformance; scores ≈ 0 suggest chance performance.

A potential concern for the evaluation of the FCT's validity is that the cut-off to define 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 psychology. 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 that simple. 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 2). The empirical malingerer sample (top) is non-normally 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 false positive rate, 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 that 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 only 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. That the non-normal nature of the malingerer distribution is unlikely to be caused by sampling error is corroborated by the prevalence of strategy levels. The detection accuracy using conservative cut-offs, such as 1% or 5% feature a sensitivity close to the prevalence of level 1 strategies. This is not surprising, as the underperformance criterion has almost perfect detection accuracy for this criterion.

From the non-normal nature of the malingerer sample follows that 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 prevalence of a level 1 strategy in 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 theoretically 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.

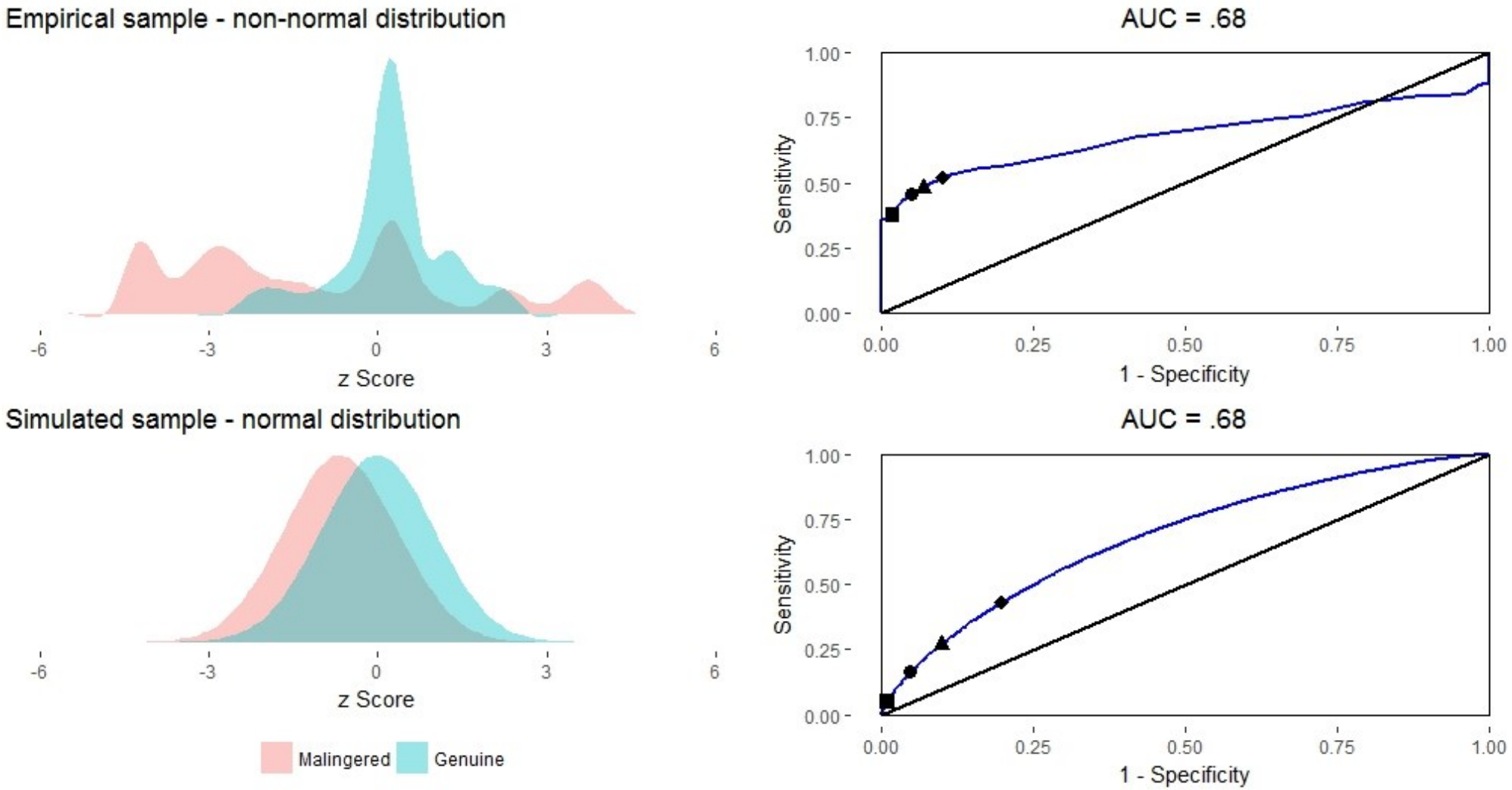

Figure 2 Distribution of z-scores for the empirical (top) and simulated (bottom) samples of malingered and genuine test performance. The left side displays the distribution overlap and the right side the Receiver Operating Curve (ROC) for each sample. Data for the simulated was generated using the effect 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%

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' test scores 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 incident rates of reported strategies. However, it affected the quality of the strategy execution. A considerable proportion of examinees reporting to choose a mixture of correct and incorrect answers still produced test scores considered below chance performance under time pressure. 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 the beliefs malingerers form, similar to Chapter 2, to misdirect malingerers in choosing 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 and 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' 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 (but not the actual) 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 is worse.

Detection accuracy can also be increased using criteria for level 1 and level 2 strategies in conjunction through 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. Malingering 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.

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. Similarly, Chapter 2 utilised the unidirectional 5% cut-off score in conjunction with absolute scores. Therefore, the single cut-off detection accuracy actually corresponds to a 10% cut-off point in terms of specificity. Still, the observed detection accuracy is in line with the data from other experiments using the same cut-off.

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 type of participants. 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. In the absence of empirical evidence, one would assume higher education to be associated with an increased likelihood to see through the FCT's rationale. Based on this assumption, detection accuracy estimates of student samples should be more conservative.

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 to ≈ 5% of malingerers. 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 marginally better than estimated.

Test construction and practical application

Besides these general limitations that apply to all FCTs mentioned here, further case specific limitations in terms of constructing a FCT must be considered. In particular, 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 potential 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 limited (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 have remembered the information probed in the test. It can be hard for the investigator to correctly estimate what pieces of information a malingerer would remember and including trials with unremembered information only reduces the

difference between malingered and genuine performance. In contrast, the maximum test size in cases of cognitive deficits is unbound. 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 pre-requisite 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 between malingerers and genuine impairment. Furthermore, detection accuracy of the response bias in Chapter 5 increased linearly with test size, which suggests that even better detection accuracies can 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 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.

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. In light of the multiple hurdle approach mentioned earlier, the FCT would serve as a first hurdle. Failing the FCT would qualify the examinee for further assessment by follow-up procedures. Here, instead of increasing the detection accuracy of the test, the FCT is used to filter out unlikely suspects in order to reduce the costs incurred by the follow-up procedure. 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.

Challenges and future directions

Several aspects of Forced Choice Testing require further attention. One of the core assumptions of the FCT is that answer alternatives pairs featured in the trials are equally plausible (Doob & Kirschenbaum, 1973). If they are not, genuine performance will trend in the direction of the bias, i.e. genuine performance will lead to lower test scores if the correct answer alternative is less likely to be selected. A consequence of biased answer alternative pairs is that the interpretation of the test score, which is evaluated in accordance with the assumed chance performance level, becomes less accurate. 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, 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 criterion 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, & Powel, 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, potential criteria for malingering included only the examinees' choices in the FCT. Interpreting the choosing pattern alone is challenging due to the non-normal distribution of malingerers' response strategies. 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.

One way measures, such as mouse dynamics, could be used is to verify the conclusions drawn from the underperformance criterion or runs test. That is, when a test score passes the underperformance threshold, the auxiliary measure is consulted to 'verify' the response pattern. Specifically, a test scores that falls below chance performance could only be considered as malingered performance when the mouse movement indicates a consistent trajectory in favour of the correct answer during trials the incorrect answer was selected. This kind of movement pattern would only be expected to be present in malingerers and therefore, could be used to identify cases when a genuinely impaired examinee produces a test score below chance performance through guessing. Consequently, this approach would reduce the number of false positives or promote the use of more liberal decision thresholds, hence increasing sensitivity, while limiting the increase in false positives.

A final concern is research on the influence a FCT exerts on subsequent aspects of criminal and clinical investigations has been neglected. 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 FCT filler trials that is trials with no correct answer alternative, could be generated that imply 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.

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 sample, 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 while retaining the reduction in false positives. 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, and, 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 is problematic for additional criteria. 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.

References

Berthelson, L., Mulchan, S.S., Odland, A.P., Miller, L.J., & Mittenberg, W. (2013). False positive diagnosis of malingering due to the use of multiple effort tests. *Brain Injury, 27,* 909 – 916. *DOI: 10.3109/02699052.2013.793400*

Bianchini, K.J., Mathias, C.W., & Greve, K.W. (2001). Symptom validity testing: A critical review. *The Clinical Neuropsychologist, 15(1), 19-45. DOI:10.1076/clin.15.1.19.1907*

Binder, L.M., Larrabee, G.J., & Millis, S.R. (2014). Intent to fail: Significance testing of forced choice test results. *The Clinical Neuropsychologist,* 28(8), 1366 – 1375. DOI: 10.1080/13854046.2014.978383

Bush, S., Heilbronner, R.L., & Ruff, R.M. (2014). Psychological assessment of symptom and performance validity, response bias, and malingering: Official position of the association for scientific advancement in psychological injury and law. *Psychological Injury and Law, 7,* 197 – 205. *DOI: 10.1007/s12207-014-9198-7*

Carmerer, C.F., Ho, T., & Chong, J. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics, 119(3), 861-898. DOI: 10.1162/0033553041502225*

Chafetz, M., & Underhill, J. (2013). Estimated costs of malingered disability. *Archives of Clinical Neuropsychology, 28,* 633 – 639. *DOI: 10.1093/arclin/act038*

Chafetz, M.D., Williams, M.A., Ben-Porath, Y.S., Bianchini, K.J., Boone, K.B., Kirkwood, M.W., Larrabee, G.J., & Ord, J.S. (2015). Official position of the American academy of clinical neuropsychology social security administration policy on validity testing: Guidance and recommendations for change. *The Clinical Neuropsychologist, 29,* 723 – 740. *DOI: 10.1080/13854046.2015.1099738*

Cima, M., Nijman, H., Merckelbach, H., Kremer, K., & Hollnack, S. (2004). Claims of crime-related amnesia in forensic patients. *International Journal of Law and Psychiatry, 27,* 215 – 221. *DOI: 10.1016/j.ijlp.2004.03.007*

Color blindness (n.d.) In *Wikipedia.* Retrieved June 1st 2018, from https://en.wikipedia.org/wiki/Color_blindness#Red%E2%80%93green_color_blindness

Davis, J.J. (2018). Performance validity in older adults: Observed versus predicted false positive rates in relation to number of tests administered. *Journal of Clinical and Experimental Neuropsychology, 40,* 1013 – 1021. *DOI: 10.1080/13803395.2018.1472221*

Denney, R.L. (1996). Symptom validity testing of remote memory in a criminal forensic setting. *Archives of Clinical Neuropsychology, 11(7),* 589-603. *DOI: 10.1093/arclin/11.7.589*

Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups – partial remembering. *Journal of Police Science and Administration, 1, 287-293.*

Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological Review, 87(3),* 215 – 251.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgement. *Psychological Review, 104(2),* 301-318.

Fechner, Gustav Theodor (1889). *Elemente der Psychophysik* (2 Volumes) (2nd ed.). Leipzig: Breitkopf & Härtel. Vol 2.

Frederick, R.I., Crosby, R.D., & Wynkoop, T.F. (2000). Performance curve classification of invalid responding on the validity indicator profile. *Archives of Clinical Neuropsychology, 15,* 281 – 300.

Frederick, R.I., & Crosby, R.D. (2000). Development and validation of the validity indicator profile. *Law and Human Behavior, 24,* 59 -82.

Frederick, R.I., & Foster, H.G. (1991). Multiple measures of malingering on a forced-choice test of cognitive ability. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3,* 596 – 602.

Giger, P., Merten, T., Merckelbach, H., & Oswald, M. (2010). Detection of feigned crime-related amnesia: A multi-method approach. *Journal of Forensic Psychology Practice, 10, 440-463. DOI: 10.1080/15228932.2010.489875*

Gudjonsson, G.H., & Shackleton, H. (1986). The pattern of scores on raven's matrices during 'faking bad' and 'non-faking' performance. *British Journal of Clinical Psychology, 25,* 35 – 41.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143, 29 – 36.*

Hancock, Peter. Psychological image collection at sterling (PICS). University of Stirling, Stirling. 30 October 2014. http://pics.psych.stir.ac.uk

Hiscock, M., & Hiscock, S. (1989). Refining the forced-choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology, 11*, 967 – 974.

Heilbronner, R.L., Sweet, J.J., Morgan, J.E., Larrabee, G.J., Millis, S.R., & Conference Participants (2009). American academy of clinical neuropsychology consensus

conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23,* 1093 – 1129. *DOI: 10.1080/13854040903155063*

Hiscock, M., & Hiscock, C.K. (1989). Refining the forced-choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology, 11(6),* 967 – 974.

Ishihara Color Blindness Test. (2011). Retrieved June 1st 2018 from http://ishiharatest.blogspot.nl/

Jelicic, M., Ceunen, E., Peters, M.J.V., & Merckelbach, H. (2011). Detecting coached feigning using the test of memory malingering (TOMM) and the structured inventory of malingered symptomatology (SIMS). *Journal of Clinical Psychology, 67,* 850 – 855. *DOI: 10.1002/jclp.20805*

Jelicic, M., Hessels, A., & Merckelbach, H. (2006). Detection of feigned psychosis with the structured inventory of malingered symptomatology (SIMS): A study of coached and uncoached simulators. *Journal of Psychopathology and Behavioural Assessment, 28,* 19 – 22. *DOI: 10.1007/s10862-006-4535-0*

Jelicic, M., Merckelbach, H., & van Bergen, S. (2004). Symptom validity testing of feigned amnesia for a mock crime. *Archives of Clinical Neuropsychology, 19, 525-531. DOI: 10.1016/j.acn.2003.07.004*

Klapproth, F. (2008). Time and decision making in humans. *Cognitive, Affective, & Behavioural Neuroscience, 8,* 509 – 524. *DOI: 10.3758/CABN.8.4.509*

Kuhn, G., Caffaratti, H. A., Teszka, R., & Rensink, R. A. (2014). A psychology-based taxonomy of misdirection. *Frontiers in Psychology, 5,* 1–14. *DOI:10.3389/fpsyg.2014.01392.*

Larrabee, G.J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society, 18, 625 – 631. DOI: 10.1017/S1355617712000240*

Larrabee, G.J. (2014). False-positive rates associated with the use of multiple performance and symptom validity tests. *Archives of Clinical Neuropsychology, 29,* 364 – 373. *DOI: 10.1093/arclin/acu019*

Larrabee, G.J. (2015). The multiple validities of neuropsychological assessment. *American Psychologist, 70, 779 – 788. DOI: 10.1037/10039835*

Meijer, E.H., Klein Selle, N., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the concealed information test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology, 51,* 879 – 904. *DOI: 10.1111/psyp.12239*

Meijer, E.H., Smulders, F.T., Johnston, J.E., & Merckelbach, H. (2007). Combining skin conductance and forced choice in the detection of concealed information. *Psychophysiology, 44,* 814-822. *DOI: 10.1111/j.1469-8986.2007.00543.x*

Merckelbach, H., Hauer, B., & Rassin, E. (2002). Symptom validity testing of feigned dissociative amnesia: A simulation study. *Psychology, Crime, & Law, 8, 311-318. DOI: 10.1080/1068316021000054256*

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*

Mittenberg, W., Patton, C., Canyock, E.M., & Condit, D.C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 8,* 1094 – 1102. *DOI: 10.1076/jcen.24.8.1094.8379*

Nickerson, R.S. (2002). The production and perception of randomness. *Psychological Review, 109(2),* 330-357. *DOI: 10.1037//0033-295X.109.2.330*

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231 – 259.

Orthey, R., Vrij, A., Leal, S., & Blank, H. (2017). Strategy and misdirection in forced choice memory performance testing in deception detection. *Applied Cognitive Psychology, 31(2),* 139-145. *DOI: 10.1002/acp.3310*

Orthey, R., Vrij, A., Meijer, E.H., Leal, S., & Blank, H. (2018). *Applied Cognitive Psychology, 32,* 1 – 8. *DOI: 10.1002/acp.3443*

Pankratz, L. (1983). A new technique for the assessment and modification of feigned memory deficit. *Perceptual and Motor Skills, 57,* 367-372.

Pankratz, L., Fausti, S.A., & Peed, S. (1975). A forced-choice technique to evaluate deafness in the hysterical or malingering patient. *Journal of Consulting and Clinical Psychology, 43(3), 421-422. DOI: 10.1037/h0076722*

Plass, J.L., Moreno, R., & Brunken, R. (2010). *Cognitive Load Theory.* Cambridge University Press; New York, NY

Podlesney, J.A. (2003). A paucity of operable case facts restricts applicability of the guilty knowledge technique in FBI criminal polygraph examinations. *Forensic Science Communications, 5,* Retrieved November, 29, 2017, from https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/july2003/podlesny.html

Rosen, G.M., Phillips, W.R. (2004) A cautionary lesson from simulated patients. *The Journal of the American Academy of Psychiatry and the Law, 32,* 132 – 133.

Schwarz, N. (1999). Self reports. How the questions shape the answers. *American Psychologist, 54(2),* 93-105.

Shaw, D. J., Vrij, A., Mann, S., Leal, S., & Hillman, J. (2012). The guilty adjustment: Response trends on the symptom validity test. *Legal and Criminological Psychology. DOI: 10.1111/j.2044-8333.2012.02070.x*

Siegel, S. (1956). *Nonparametric statistics.* New York: McGraw-Hill.

Siegel, S., & Castellan, N.J. (1988). *Nonparametric statistics for the behavioural sciences.* New York: McGraw-Hill.

Simon, H. A. (1955). A behavioural model of rational choice. *Quarterly Journal of Economics, 69,* 99 – 118.

Slick, D.J., Sherman, E.M.S., & Iverson, G.L. (1999). Diagnostic criteria for malingered neurocognitive hysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 13,* 545 – 561. *DOI: 10.1076/1385-4046(199911)13:4;1-Y;FT545*

Tanner, W.P., & Swets, J.A. (1954). A decision-making theory of visual detection. *Psychological Review, 61,* 401 – 409. *DOI: 10.1037/h0058700*

Van Impelen, A., Jelicic, M., Otgaar, H., & Merckelbach, H. (2017). Detecting feigned cognitive impairment with schretlen's malingering scale vocabulary and abstraction test. *European Journal of Psychological Assessment. DOI: 10.1027/1015-5759/a000438*

Van Oorsouw, K., & Merckelbach, H. (2006).Simulating amnesia and memories of a mock crime. *Psychology, Crime & Law, 12,* 261 – 271. *DOI: 10.1080/10683160500224477*

Van Oorsouw, K., & Merckelbach, H. (2010). Detecting malingered memory problems in the civil and criminal arena. *Legal and Criminological Psychology, 15, 97 – 114. DOI: 10.1348/135532509X451304*

Verschuere, B., Meijer, E., & Crombez, G. (2008). Symptom validity testing for the detection of simulated amnesia: Not robust to coaching. *Psychology, Crime, & Law, 14(6),* 523-528. *DOI: 10.1080/10683160801955183*

Wagenaar, W.A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin, 77,* 65-72.

Zvi, L., Nachson, I., & Elaad, E. (2012). Effects of coping and cooperative instructions on guilty and informed innocents' physiological responses to concealed information. *International Journal of Psychophysiology, 84*, 140 – 148.

Zvi, L., Nachson, I., & Elaad, E. (2015). Effects of perceived efficacy and prospect of success on detection in the guilty action test. *International Journal of Psychophysiology, 95,* 35 – 45.

Summary

Malingered cognitive impairment, such as loss of hearing, or malingered loss of memory for specific events, for example a crime, can be detected with the Forced Choice Test (FCT). This test features a series of binary multiple choice trials and forces the examinee to make a choice on each trial. Genuinely impaired performance is defined as guessing, producing test scores within levels of chance. In contrast, malingered performance is defined as the intentional avoidance of correct answers, producing test scores worse than chance performance. The aim of this thesis was twofold. First, to explore the FCT's optimal detection accuracy, and second, to establish a model that describes the various counterstrategies malingerers use to defeat this test.

Chapter 2 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 we define three strategy levels for malingerers in the FCT ranging from 0 to 2. 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. We subjected liars and truth tellers 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 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 we introduced time pressure to the FCT paradigm, by forcing examinees with and without concealed information to select an answer alternative within two seconds. We compared this paradigm 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 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, we attempted to force examinees to choose between a randomizing pattern that 'looks' random and a randomizing pattern that produces a test score within chance levels. To do so we alternate the position of the correct answer alternative (left or right) 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. We expected that coached examinees would 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 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, we aimed at introducing a manipulation, independent of the actual task (here to discriminate red and green), that elicits a systematic preference. We varied the perceived difficulty of the trials 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, we instructed examinees 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. We re-examined the validity of the 'runs-test', 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.

Finally, Chapter 6 will feature an evaluation of the three strategy levels and their correspondence to malingerers' test behaviour as well as a reassessment of the FCT's detection accuracy. Pathways to increase detection are discussed and two fields of application, malingered sensory dysfunction and malingered loss of memory, are differentiated in terms of test construction and choice of criteria. The chapter closes with a reflection on experimental limitations and future challenges.

VALORIZATION ADDENDUM

Innovation

The novel contributions of this thesis lie in the conceptualization of malingerers' response strategies to defeat the Forced Choice Test and the development of additional criteria based on those strategies. Originally, malingerers were believed to simply avoid selecting correct answers leading to test scores lower than expected by chance performance. The problem here is that this response strategy applies to only half of the malingerers in a typical experiment. The presence of other response strategies has been mentioned before, but has never been formalized. This thesis extends the original model of malingerers' response strategies to incorporate these additional response pattern. In particular it defines intentional randomization and endorsement of correct answers as potential response strategies and a reflection on the relationship between the detection accuracy of the Forced Choice Test and the prevalence of these three subgroups.

Especially innovative is the introduction of two conditions under which the runs test has diagnostic value. The runs test measures the alternations between correct and incorrect answer alternatives and has been proposed as a suitable measure for random responding (e.g. Verschuere, Meijer, & Crombez, 2008). However, previous experiments failed to identify malingerers using this test (see Verschuere et al., 2008; Jelicic, Merckelbach, & van Bergen, 2004). In this thesis, it was demonstrated that the runs test either requires a large test size (see Chapter 5) or a specific change in the Forced Choice Test paradigm (see Chapter 4). These simple manipulations increased the diagnostic validity of the runs test, and can be easily implemented in practice by the relevant target groups such as neuropsychologists.

Another innovation in this thesis is the proposed response bias criterion based on perceived difficulty for randomising behaviour. The idea is simple, examinees who understand that their final test score has to fall within levels of chance know that they must select correct and incorrect answers. This presents malingerers with the challenge of deciding when and under what circumstances to select the correct answer alternatives. The new criterion is based on the idea that the process that leads to a test score within chance performance can be influenced in order to become systematic and therefore distinguishable from actual chance performance. In Chapter 5 malingerers produced six times as much test patterns outside chance performance than would be expected by a truly random process and the likelihood of their response patterns had a good diagnostic value.

Relevance

The data and conclusions from this thesis bear significance to both academics and practitioners. Academics can profit from the extended definition of malingerers' response strategies and the proposed underlying mechanism. A direct benefit of the distinction of various response strategies is that more specific hypothesis can be generated and that statistical tests can be conducted per subgroup, as the non-normal distribution of the combined sample limits the credibility of the typically employed analyses. Furthermore, the data presented here suggests that future research should focus on developing new criteria sensitive to randomization behaviour instead of the traditional underperformance criterion. Hence, this thesis provides academics with an overview of malingerers' behaviour and acts as a foundation for the development of new criteria and manipulations to increase the detection accuracy of the Forced Choice Test.

For practitioners, this thesis demonstrates easy to implement new criteria sensitive to intentional randomisation. These criteria can be used to increase the detection accuracy of the Forced Choice Test, because the largest subgroup of malingerers follows a randomisation strategy and hence avoids detection trough the traditional underperformance criterion. Furthermore, the reflection on the non-normal distribution of malingerers' test responses and its relation to single cut-off point detection accuracy can aid practitioners in better determining what test results should be considered as malingered performance. In particular, the data presented here calls the 5% cut off point that is traditionally applied into question. As suggested earlier the choice for this cut off was likely convention in the field rather than empirical observation. Consequently, practitioners are invited to reconsider this choice of cut off. For example, a much more conservative 1% cut off would yield a similar detection accuracy for malingered performance while greatly reducing the false positive rate.

Future Directions

Future directions must prioritize the applicability of the Forced Choice Test. As it stands, the test can only be constructed for cases that provide large amounts of information that the examinee must remember. Although the recommended minimum test length varies from 12 (Van Oorsouw, & Merckelbach, 2010) to 25 (Denney, 1996), many crimes/events do not fulfil this requirement and therefore the use of the Forced Choice Test is limited to few selected cases.

Acknowledgements

The last four years passed in the blink of an eye and felt like a caffeine fuelled fever dream. I lived in over three countries, met more people in this time than I had known before in total, and was presented with many challenges. If I know one thing for certain, it is that none of this would have been possible without you, Brigitte, Jörg, and Max. Thank you for always being there and supporting me.

In this project, I was fortunate enough to conduct research at two universities in Europe. I am grateful to Maastricht University and the University of Portsmouth for giving me the opportunity and freedom to conduct research. I am also grateful to my supervisors, Aldert, Ewout, Sharon, and Hartmut, who not only tolerated but encouraged my research ideas. Thank you for your guidance, patience, and encouragement.

Furthermore, I wish to express my gratitude to my friends close and afar. You are many and spread over the entire northern and a bit of the southern hemisphere. To name but a few and in random sequence: Liam, Nicola, Irena, Jakob, Hannah, Kathy, Adam, Char, Sarah, Ivan, Nael, Phillip, Alexander, Sergii, Brie, Ali, and many more… Thank you for your advice, help, companionship and the great times we had together. I hope will meet again soon.

Finally, Wei, without your help and support I would have never managed to overcome the last year with all its challenges. Thank you for always being there and supporting me.

About the Author

Robin Orthey was born on 15th of June 1990 in Lüdenscheid, Germany. He completed a Bachelor degree (cum laude) in psychology as well as disciplinary honours program (cum laude) at the Radboud University Nijmegen. This was followed by a master degree (cum laude) in Psychology & Law at Maastricht University. In 2014 he pursued a PhD in psychology at the University of Portsmouth and Maastricht University by investigating the human behaviour in the Forced Choice Test. During this project he was placed as a visiting researcher in Bar-Ilan University, Israel, and Fukuyama University, Japan, to foster and establish international connections. So far, this project has resulted in two publications in peer reviewed journals and two presentations at international conferences.

DISSEMINATION

Publications

Orthey, R., Vrij, A., Leal, S., & Blank, H. (2017). Strategy and misdirection in forced choice memory performance testing in deception detection. *Applied Cognitive Psychology, 31, 139-145. DOI: 10.1002/acp.3310*

Orthey, R., Vrij, A., Meijer, E., Leal, S., & Blank, H. (2018). Resistance to coaching in forced choice testing. *Applied Cognitive Psychology, 32, 1 - 8. DOI: 10.1002/acp.3443*

Orthey, R., Vrij, A., Meijer, E., Leal, S., & Blank, H. (2019). Eliciting Response Bias Within Forced Choice Tests to Detect Random Responders, *Nature Scientific Reports*

Orthey, R., Palena, N., Vrij, A., Meijer, E., Leal, S., Blank, H., & Caso L. (pending minor revision). Effects of Time Pressure on Strategy Selection and Strategy Execution in Forced Choice Tests. *Applied Cognitive Psychology*

Boskovic, I., Dibbets, P., Bogaard, G., Hope, L., Jelicic, M., & **Orthey, R**. (2019). Verify the scene, report the symptoms: Testing the verifiability approach and SRSI in the detection of fabricated PTSD claims. *Legal and Criminological Psychology. DOI: 10.1111/lcrp.12149*

Palena, N., Caso, L., Vrij, A., & **Orthey, R**. (2018). Detecting deception through small talk and comparable truth baseline. *Journal of Investigative Psychology and Offender profiling, 1 – 9.*

McCarthy, R. J., … **Orthey, R.**, …, Yildiz, E. (2018). Registered Replication Report: Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science, 1(3),* 299 - 317. *DOI: 10.1177/2515245918781032*

Verschuere, B., …, **Orthey, R.**, …, Yildiz, E. (2018). Registered Replication Report: Mazar, N., Amir, O., & Ariely, D. (2008). *Advances in Methods and Practices in Psychological Science, 1(3),* 321 - 336. *DOI:10.1177/2515245918777487*

Conference Presentations

2018 – EAPL 2018

- Oral presentation: *Forced Choice Testing with limited amount of information*

2017 – CIT meeting in Fukuyama: Verification of new indices on CIT

- Oral presentation: *Detecting concealed information using Forced Choice Testing*, August 26 & 27th

2017 – National Research Institute for Police Science (Japan)

- Oral presentation: *Detecting concealed information using Forced Choice Testing*

2017 – Kwansei Gakuin University CAPS meeting:

- Oral presentation: *Detecting deception with the forced choice paradigm*, August 1st

2017 – EAPL 2017

- Oral presentation: Resistance to Countermeasures in Forced Choice Testing