

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26

Resistance to Coaching in Forced Choice Testing

Robin Orthey^{1,2}

Aldert Vrij¹

Ewout Meijer²

Sharon Leal¹

Hartmut Blank¹

¹ University of Portsmouth

² Maastricht University

1 **Abstract**

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

18

19

20

21

22

23

24

25

26

27

28

29

30

1

2

Resistance to Coaching in Forced Choice Testing

3

4 Forced choice testing (FCT) has been used as a test to detect malingering of sensory
5 impairment (Pankratz, Fausti, & Peed, 1975). More recently, it's use has been extended to
6 detect cases of faked memory loss (e.g., Denney, 1996; Hiscock & Hiscock, 1989; Pankratz,
7 1983; Van Oorsouw, & Merckelbach, 2010) and concealed information (e.g., Giger, Merten,
8 Merckelbach, & Oswald, 2010; Meijer, Smulders, Johnston, & Merckelbach, 2007; Orthey,
9 Vrij, Leal, & Blank, 2017; Shaw, Vrij, Mann, Leal & Hillman, 2012), from which guilty
10 knowledge can be inferred. In the case of concealed information detection, a typical test
11 works as follows: A suspect is presented with a series of questions about the crime. With
12 each question, two equally plausible answer alternatives are presented; a correct and an
13 incorrect one. For example a question such as "What was the murder weapon" could be
14 accompanied with two answer alternatives such as "gun" and "knife". Suspects are instructed
15 to select the correct answer, or guess if they don't know. Innocent suspects – who have no
16 knowledge of the correct answers – will have to guess on each trial, and thereby choose
17 correct answer alternatives as predicted by chance. Guilty suspects, in contrast, know which
18 of the two alternatives is correct. To conceal this guilty knowledge, they are inclined to
19 purposefully select the incorrect answers, leading to underperformance, i.e., the frequency
20 with which the correct option is chosen is below chance level. Consequently, hidden
21 knowledge is inferred from underperformance.

21

22 Previous studies have shown that FCTs have good detection rates for innocent
23 examinees, specificity. However, the detection rate for guilty examinees, sensitivity, is
24 modest at best. More specifically, with a specificity ranging around 95%, sensitivity ranges
25 from 40% to 65% (Giger et al., 2010; Jelicic, Merckelbach, & van Bergen, 2004; Meijer et al.,
26 2007; Merckelbach, Hauer, & Rassin, 2002; Shaw et al., 2012). These validity estimates are,
27 however, for participants who are unfamiliar with the test's underlying rationale. Verschuere,
28 Meijer, and Crombez (2008) showed that sensitivity is reduced considerably when
29 participants have been informed about this rationale (i.e. coached). These authors coached
30 half of their participants, and then submitted both naïve and coached participants to a forced
31 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

1 correct items selected) as the criterion. Consequently, the authors conclude that forced choice
2 performance testing is not resistant to coaching.

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

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

1 Using a level 2 strategy means that examinees will attempt to produce a random
2 sequences of correct and incorrect answers to pass the test. Yet, the correct total criterion is
3 not the only criterion of randomness. Another criterion is the alternation rate. For example the
4 sequence of CORRECT CORRECT CORRECT INCORRECT INCORRECT INCORRECT
5 contains one alternation. The sequence of CORRECT INCORRECT CORRECT
6 INCORRECT CORRECT INCORRECT contains 5 alternations. Innocent examinees
7 alternate between correct and incorrect answers on subsequent trials at a rate of 50%. Yet it is
8 not the case for guilty examinees. There is strong evidence suggesting that humans cannot
9 properly reproduce randomness. When asked to generate a random response pattern, humans
10 were found to utilize higher alternation rates than expected from true randomness (Nickerson,
11 2002; Wagenaar, 1972). Multiple estimates suggest that human random responding features
12 an alternation rate of 60% as opposed to randomness's alternation rate of 50% (see Falk &
13 Konold, 1997). In other words, an attempted random mixture of correct and incorrect answers
14 can be expected to exhibit more alternations than a genuine random response pattern.

15 Indeed, the number of alternations between correct and incorrect has been used to
16 detect coached participants, but with limited success. Verschuere et al. (2008) only identified
17 21% coached liars. Similarly, Jelicic et al., (2004) – tested the number of alternations in those
18 participants who indicated randomization as their strategy. In their sample not a single liar
19 was identified using this test.

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

27 In real life, including many items in forced choice performance deception detection
28 tests may not always be feasible. The event may, for example not have enough details the
29 investigators can verify and are exclusively known to the perpetrator (Podlesney, 1983). If
30 constructing large tests is not possible, another way to enhance detection accuracy is needed.

31 In this experiment we attempted to increase the diagnostic accuracy of the FCT
32 procedure without requiring additional questions. Traditionally, each question in a forced

1 choice test is presented with two answer alternatives. The position of the correct answer
2 alternative (e.g., left or right) is determined randomly for each trial. In the current experiment,
3 we alternate the position of the correct answer alternative between trials. On the first trial the
4 horizontal position of the correct answer alternative would be determined randomly, for
5 example on the right. On every subsequent trial the correct answer alternative would be
6 presented on the opposite side of the previous trial. This way of presenting the answer
7 alternatives allows for two types of randomized response patterns: Guilty examinees can
8 randomize horizontally, alternating between left and right answer alternatives (which will
9 look like a random response pattern), or between correct and incorrect answer alternatives
10 (which produces a total score that falls within chance performance). In our design,
11 correct/incorrect and horizontal alternations become negatively correlated. A high number of
12 correct/incorrect alternations is associated with a low number of horizontal alternations and
13 vice versa (e.g., always choosing the option presented on the left results in the maximum
14 number of correct/incorrect alternations as well as the lowest number of horizontal
15 alternations). Our idea behind this manipulation is as follows: innocent participants – whether
16 naïve or coached – are unaware of which of the answer alternatives is correct, and will choose
17 to randomize horizontally. As a consequence they will show a high number of horizontal
18 alternations, corresponding to a low number of correct/incorrect alternations. Coached guilty
19 participants are expected to employ level 2 strategies and are faced with having to choose
20 between producing a sequence that looks ‘random’ (high frequency of horizontal alternations)
21 or producing a sequence where the correct total criterion falls within chance levels. Being
22 aware of the underlying rationale of FCT will likely result in a high number of
23 correct/incorrect alternations. In naïve guilty examinees we expect all strategy levels to occur
24 naturally with prevalences similar to Orthey et al. (2017), and that different criteria can detect
25 different strategies. So the total score criterion will detect the examinees who employ level 1
26 strategies, while the number of runs criterion will detect examinees who employ level 2
27 strategies.

28 Specifically, in the current study we investigated two questions:

- 29 i) What is the effect of coaching on the strategies guilty and innocent participants
30 select?
- 31 ii) Can correct/incorrect alternations that are correlated with horizontal
32 positioning discriminate guilty from innocent participants in cases of
33 coaching?

1 examinees who had experienced the intelligence scenario (henceforward referred to as guilty
2 examinees), were instructed to lie and to convince the police that they had never been in the
3 terrorist's apartment. Examinees who had experienced the real estate scenario (henceforward
4 referred to as innocent examinees), were informed that they never had been to the terrorist's
5 apartment and that they were falsely accused. They were told that it was their task to convince
6 the investigators that they had no knowledge of the terrorist apartment. Then examinees were
7 randomly divided into a coached (N = 52) and naïve condition (N = 52), evenly split over the
8 two VR scenarios. Coached examinees were provided with an advice from their attorney
9 warning them about the mechanisms of the lie detection test (naïve examinees received no
10 such information and directly moved on to the next part). Coached examinees received the
11 following information:

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

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

29 After completing the FCT all examinees were informed that the lie detection test was
30 over and that they should answer the post-test questions honestly. First, they received two
31 open questions, 'What did you do to appear innocent during the lie detection test?' and 'What
32 strategy did you have in mind to make the investigator believe that you were uninvolved with

1 *the terrorist?*'. Then guilty examinees received the questions and answer alternatives again
2 and had to indicate the correct answer for each question, which referred to the actual stimulus
3 encountered in the intelligence scenario. This served as a memory check. Guilty examinees
4 remembered on average 95% of the correct answers ($SD = 5.6$; worst performance = 80%).

5

6

7

8 **Forced Choice Test**

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

18 **Design and Measures**

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

27 Detection accuracy was measured in terms of sensitivity and specificity. Sensitivity
28 indicates the proportion of guilty participants correctly classified and specificity indicates the
29 proportion of innocent participants correctly classified. Sensitivity and specificity are based
30 on a specific cut off point. For the correct total the cut off was based on the theoretical binary

1 distribution as we expect innocent participants to inadvertently follow it. Sensitivity and
2 specificity were computed for the conventionally used unidirectional 5% specificity cut off, as
3 well as for 10% and 20% cut offs (e.g. Binder, Larrabee, & Millis, 2014; Van Impelen,
4 Jellicic, Otgaar, & Merckelbach, 2017).

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

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

22 Assessing the accuracy of such a two-step procedure is relevant, because level 2
23 strategies occur naturally in naïve guilty. In fact, in Orthey et al. (2017) it was the most
24 prevalent strategy, meaning that the runs-criterion could be relevant even for cases without
25 coaching. Furthermore, as seen in Orthey et al. (2017) some examinees who employed level 2
26 strategies still were detected using the total score criterion, likely because they incorrectly
27 judged how many correct items were required for the test score to still fall within chance
28 performance. Therefore, we must estimate how many cases of level 2 strategies still get
29 detected by the total score criterion, as these cases would have been detected anyway. The
30 remaining detection accuracy then indicates the incremental validity of detecting intentional
31 randomization. As sensitivity and specificity correspond to a specific cut off point they do not
32 generalize to other cut offs. Instead, the Area Under the Curve (AUC) can be used as an

1 indicator for detection accuracy independent of cut off points. It is based on the Receiver
2 Operator Characteristic (Tanner & Swets, 1954; ROC), which plots sensitivity against
3 specificity for the entire range of the continuous criterion. The AUC is the area covered by the
4 ROC. It ranges between 0 to 1 with 0.5 indicating chance performance, and a higher number
5 meaning better discrimination between guilty and innocent examinees.

6 Participants answers to the open questions about their behaviour during the test were
7 categorised into three strategy levels. Level 0 strategies represented compliance with the test
8 instructions to select the correct answers alternatives. Participants who indicated that they
9 selected answers they thought were correct or those who indicated to use no strategy were
10 assigned to this level. Level 1 strategies represented a reaction to the test instructions.
11 Participants who said they avoided correct answers on purpose or controlled their demeanour
12 while selecting answers were assigned to this level. Level 2 represented patterns that
13 purposefully included correct and incorrect answers. Participants who said they imitated
14 responses patterns they believe people ignorant of the crime information would produce, or
15 said they selected answers that seem obvious (either correct or incorrect), or indicated
16 purposefully randomising between correct and incorrect answers were assigned to this level.
17 Two blind and independent raters categorised the responses according to examples within
18 each strategy level as specified in Orthey et al. (2017). Inter rater reliability was high (89%
19 absolute agreement). Responses that did not fit any category were omitted from the analysis
20 (1 participant).

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

30

31

Results

1 **Strategies**

2 **TABLE 1 HERE**

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

10 A chi-square test was performed and we found a relationship between coaching and
11 the used strategy level for guilty examinees, $\chi^2(2, N = 51) = 16.32, p < .001$. Coached guilty
12 examinees were more likely to exhibit a level 2 strategy than naïve guilty examinees. A closer
13 look at the data revealed that the entire sample of coached guilty examinees used a level 2
14 strategy, whereas the naïve guilty examinee sample consisted out a number of level 0, 1, and
15 2 strategies ($M = 1.44, SD = 0.65$). This supports Hypothesis 1.

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

23

24 **Detection Accuracy**

25 **TABLE 2 HERE**

26

FIGURE 1 HERE

1 First we examined the correct total criterion. In the naïve condition a low correct total
2 differentiated guilty from innocent examinees better than chance¹, $AUC = .69$, $p = .020$, $CI =$
3 $[.53 .86]$. In the coaching condition the correct total did not distinguish guilty from innocent
4 examinees better than chance, $AUC = .53$, $p = .742$, $CI = [.37 .69]$. Similarly, when using the
5 conventionally used unidirectional decision cut off of 5%, we found a 48% sensitivity and a
6 92% specificity in naïve guilty examinees. Using a 10% cut off sensitivity rose to 56% while
7 specificity remained the same at 92.3%. At the 20% cut off sensitivity was 64% with a
8 specificity of 88.5%. When coached, the sensitivity dropped to 7.7% with a 100% specificity
9 at the 5% cut off. At the 10% cut off sensitivity remained at 7.7%, but specificity declined to
10 92.3%. At the 20% cut off sensitivity was 11.5% with a specificity of 88.5%. This suggested a
11 sharp decline in detection accuracy for the correct total criterion in case of coaching, which
12 supports Hypothesis 2.

13 Next we examined the runs criterion. In the naïve condition, a high number of
14 alternations resulted in worse general detection accuracy than chance¹, $AUC = .26$, $p = .008$,
15 $CI = [.14 .43]$. However, in the coaching condition the number of runs differentiated guilty
16 from innocent examinees significantly better than chance performance, $AUC = .69$, $p = .018$,
17 $CI = [.55 .84]$. We examined the detection accuracy for multiple suggested single cut offs and
18 used the unidirectional cut offs of 5%, 10%, and 20%. In the naïve condition, the runs
19 criterion featured a 0% sensitivity at the 5% cut off, which rose to 8% for the 10% and 20%
20 cut off. Specificity was highest for the 5% and 10% cut offs with 92.31%. At the 20% cut off
21 it declined to 80.71%. In the coaching condition, the 5% cut off featured a 7.69% sensitivity
22 and 100% specificity. At the 10% cut off sensitivity increased to 34.62%, but specificity
23 declined to 96.15%. At the 20% cut off sensitivity was 57.69% and specificity was at 69.23%.
24 Thus, for both conditions the best sensitivity/specificity ratio was found at the 10% cut off. In
25 any case the AUCs indicate that number of runs criterion was able to detect coached guilty
26 examinees, supporting Hypothesis 3.

27 Additionally, we expressed the difference between guilty and innocent examinees for
28 the correct total and runs criterion in terms of their effect size *Cohen's d*. However, this
29 indicator was only computed for the coaching condition, as only in this condition the entire

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.

1 guilty sample utilized the same strategy level and was therefore assumed to be normally
2 distributed. We found no effect for the correct total criterion (Cohen's $d = -0.02$), as the
3 coached guilty examinees ($M = -0.38, SD = 1.26$) matched the responses of coached innocent
4 examinees ($M = -0.36, SD = 0.99$). The runs criterion had a medium effect (Cohen's $d = -$
5 0.41), as coached guilty examinees ($M = -0.05, SD = 1.18$) favored alternating between
6 correct and incorrect answer alternatives, but coached innocent examinees prioritized
7 alternations between horizontal positions ($M = -0.46, SD = 0.91$).

8 **Incremental Validity**

9 **TABLE 3 HERE**

10 Finally we assessed the incremental validity of a two-step classification process. As
11 step 1 we used the correct total criterion with the conventional unidirectional cut off at
12 5%. That is, all participants whose correct total score fell within underperformance were
13 classified as guilty. As the second step the remaining sample was subjected to the runs
14 criterion using the three unidirectional cut offs 5%, 10%, and 20%. Accuracy was expressed
15 as the combined detection accuracy of steps 1 and 2. See table 3 for corresponding
16 sensitivities and specificities. The best ratio of sensitivity/specificity was found at the 10% cut
17 off. In the naïve condition, we found a sensitivity of 56% and a specificity of 84.62%. In the
18 coaching condition, sensitivity was at 42.31% with a specificity of 96.15%. Combined
19 detection accuracies indicated that sensitivity and specificity of steps 1 and 2 were additive,
20 suggesting a unique contribution from each criterion.

21 **Discussion**

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

30 Regarding the strategies examinees reported, frequencies of strategy levels in our
31 naïve condition closely matched those reported in Orthey et al. (2017). Coaching increased

1 the reported strategy level for guilty examinees and coached guilty examinees exclusively
2 reported using level 2 strategies. This is also reflected in the detection accuracy of the correct
3 total criterion per strategy level. In naïve guilty examinees, the test detected level 1 strategies
4 well, but not level 2 strategies. Similarly, detection accuracy for level 2 strategies in our
5 coaching condition was very low.

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

18 Detection accuracy in our naïve condition matched that of other experiments, as did
19 the decline in detection accuracy in our coaching condition for the correct total criterion. As
20 expected in our naïve condition we found a moderate sensitivity (48%) and good specificity
21 (92%), which matched the range of previous experiments using naïve examinees (Giger et al.,
22 2010; Jelcic et al., 2004; Meijer et al., 2007; Merckelbach et al., 2002; Orthey et al., 2017;
23 Shaw et al., 2012). In the presence of coaching sensitivity declined (8%), but specificity
24 remained high (100%), matching the findings in Verschuere et al. (2008), reinforcing their
25 conclusion that forced choice testing is not resistant to coaching when using correct total
26 criterion.

27 The AUC of the runs criterion in the naïve condition suggests below chance accuracy
28 levels. With a 10% cut off, this criterion featured a 8% sensitivity and a 92.31% specificity.
29 This poor detection accuracy is likely a consequence of the underlying abnormal strategy
30 level distribution. This criterion is geared towards detecting level 2 strategies, which made up
31 only 40% of the naïve sample. Hence sensitivity is expected to be low. Furthermore, the poor
32 AUC is explained by the substantial presence of level 1 strategies, because underperformance

1 is negatively related to the number of runs. Selecting only incorrect answers, also means not
2 switching between correct and incorrect answers, which is what the runs criterion was
3 intended to detect. Hence its' detection accuracy is poor when alone applied to all strategy
4 levels at once.

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

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

18 The combined detection accuracy of the two-step classification process with the
19 correct total criterion and alternations criterion suggests that the effects of each criterion are
20 additive. Thus, each criterion captured a unique subgroup of our guilty samples. The correct
21 total criterion was sensitive to participants using level 1 strategies (e.g. avoiding correct
22 information) and the runs criterion to those using level 2 strategies (mixture of correct and
23 incorrect answers). Consequently, the runs criterion provides incremental validity to the FCT
24 paradigm by detecting intentional randomisation either occurring naturally or as a
25 consequence of coaching.

26 The argument can be made that we coached examinees specifically regarding the
27 correct total criterion, and that similarly coaching can be extended to incorporate the runs
28 criterion as well. Nevertheless, our findings are still relevant for two reasons. First, as level 2
29 strategies also occur in naïve examinees, the runs criterion can increase the detection accuracy
30 in naïve examinees. Secondly, trying to apply countermeasures for multiple criteria at once is
31 difficult and likely taxing on cognitive resources, thus reducing the likelihood to succeed.

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

9 Another potential concern is the validity of verbal self-reports as our measure for
10 strategies. There has been considerable debate about the question how accurate self-reported
11 measures are (Nisbett, & Wilson, 1977; Ericsson, & Simon, 1980; Schwarz, 1999). The
12 concern is that human subjects may not be aware of the true reasons of their behavior and
13 when asked about it can only produce a post hoc rationalization. To address this issue we
14 specifically kept our questions focused on actual test behavior (i.e., ‘What did you do to
15 defeat the test?’ instead of ‘What was your strategy to defeat the test?’). Therefore, the impact
16 of measurement unreliability is kept to a minimum.

17 In sum, we found further support for the idea that guilty examinee’s test behavior is
18 governed by a strategy selection process based on their beliefs over the test’s mechanism. We
19 conclude that the correct total criterion is vulnerable to coaching, but coached guilty
20 examinees can be detected using our modified runs test.

21
22
23
24
25
26
27
28

References

- 1
2 Binder, L. M., Larrabee, G. J., & Millis, S. R. (2014). Intent to fail: Significance testing of
3 forced choice test results. *The Clinical Neuropsychologist*, 28(8), 1366–1375.
4 DOI:10.1080/13854046.2014.978383
- 5 Carmerer, C.F., Ho, T., & Chong, J. (2004). A cognitive hierarchy model of games. *The*
6 *Quarterly Journal of Economics*, 119(3), 861-898. DOI: 10.1162/0033553041502225
- 7 Denney, R.L. (1996). Symptom validity testing of remote memory in a criminal forensic
8 setting. *Archives of Clinical Neuropsychology*, 11(7), 589-603. DOI:
9 10.1093/arclin/11.7.589
- 10 Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups – partial remembering.
11 *Journal of Police Science and Administration*, 1, 287-293.
- 12 Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological Review*, 87(3),
13 215-251.
- 14 Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for
15 judgement. *Psychological Review*, 104(2), 301-318.
- 16 Giger, P., Merten, T., Merckelbach, H., & Oswald, M. (2010). Detection of feigned crime-
17 related amnesia: A multi-method approach. *Journal of Forensic Psychology Practice*,
18 10, 440-463. DOI: 10.1080/15228932.2010.489875
- 19 Hiscock, M., & Hiscock, C.K. (1989). Refining the forced-choice method for the detection of
20 malingering. *Journal of Clinical and Experimental Neuropsychology*, 11(6), 967 – 974.
- 21 Jelacic, M., Merckelbach, H., van Bergen, S. (2004). Symptom validity testing of feigned
22 amnesia for a mock crime. *Archives of Clinical Neuropsychology*, 19, 525-531. DOI:
23 10.1016/j.acn.2003.07.004
- 24 Meijer, E.H., Smulders, F.T., Johnston, J.E., & Merckelbach, H. (2007). Combining skin
25 conductance and forced choice in the detection of concealed information.
26 *Psychophysiology*, 44, 814-822. DOI: 10.1111/j.1469-8986.2007.00543.x

- 1 Merckelbach, H., Hauer, B., & Rassin, E. (2002). Symptom validity testing of feigned
2 dissociative amnesia: A simulation study. *Psychology, Crime, & Law*, 8, 311-318. DOI:
3 10.1080/1068316021000054256
- 4 Nickerson, R.S. (2002). The production and perception of randomness. *Psychological Review*,
5 109(2), 330-357. DOI: 10.1037//0033-295X.109.2.330
- 6 Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on
7 mental processes. *Psychological Review*, 84(3), 231-259.
- 8 Pankratz, L. (1983). A new technique for the assessment and modification of feigned memory
9 deficit. *Perceptual and Motor Skills*, 57, 367-372.
- 10 Pankratz, L., Fausti, S.A., & Peed, S. (1975). A forced-choice technique to evaluate deafness
11 in the hysterical or malingering patient. *Journal of Consulting and Clinical Psychology*,
12 43(3), 421-422. DOI: 10.1037/h0076722
- 13 Podlesney, J.A. (2003). A paucity of operable case facts restricts applicability of the guilty
14 knowledge technique in FBI criminal polygraph examinations. *Forensic Science*
15 *Communications*, 5, Retrieved November, 29, 2017, from
16 [https://archives.fbi.gov/archives/about-us/lab/forensic-science-](https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/july2003/podlesny.htm)
17 [communications/fsc/july2003/podlesny.htm](https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/july2003/podlesny.htm)
- 18 Schwarz, N. (1999). Self reports. How the questions shape the answers. *American*
19 *Psychologist*, 54(2), 93-105.
- 20 Shaw, D. J., Vrij, A., Mann, S., Leal, S., & Hillman, J. (2012). The guilty adjustment:
21 Response trends on the symptom validity test. *Legal and Criminological Psychology*.
22 DOI: 10.1111/j.2044-8333.2012.02070.x
- 23 Siegel, S. (1956). *Nonparametric statistics for the behavioural sciences*. New York: McGraw-
24 Hill.
- 25 Tanner, W.P., & Swets, J.A. (1954). A decision-making theory of visual detection.
26 *Psychological Review*, 61(6), 401-409. DOI: 10.1037/h0058700
- 27 Orthey, R., Vrij, A., Leal, S., & Blank, H. (2017). Strategy and misdirection in forced choice
28 memory performance testing in deception detection. *Applied Cognitive Psychology*,
29 31(2), 139-145. DOI: 10.1002/acp.3310

- 1 Van Impelen, A., Jelicic, M., Otgaar, H., & Merckelbach, H. (2017). Detecting feigned
2 cognitive impairment with schretlen's malingering scale vocabulary and abstraction
3 test. *European Journal of Psychological Assessment*. DOI: 10.1027/1015-
4 5759/a000438
- 5 Van Oorsouw, K., & Merckelbach, H. (2010). Detecting malingered memory problems in the
6 civil and criminal arena. *Legal and Criminological Psychology*, 15, 97 – 114. DOI:
7 10.1348/135532509X451304
- 8 Verschuere, B., Meijer, E., & Crombez, G. (2008). Symptom validity testing for the detection
9 of simulated amnesia: Not robust to coaching. *Psychology, Crime, & Law*, 14(6), 523-
10 528. DOI: 10.1080/10683160801955183
- 11 Wagenaar, W.A. (1972). Generation of random sequences by human subjects: A critical
12 survey of literature. *Psychological Bulletin*, 77, 65-72.
- 13 Zvi, L., Nachson, I., & Elaad, E. (2012). Effects of coping and cooperative instructions on
14 guilty and informed innocents' physiological responses to concealed information.
15 *International Journal of Psychophysiology*, 84, 140 – 148.
- 16 Zvi, L., Nachson, I., & Elaad, E. (2015). Effects of perceived efficacy and prospect of success
17 on detection in the guilty action test. *International Journal of Psychophysiology*, 95, 35
18 – 45.
- 19
20
21
22
23
24
25
26
27

1
2
3
4
5

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

6
7
8
9
10
11
12

Table 2. Detection accuracy for the alternations criterion

	Sensitivity			Specificity			AUC	<i>p</i>	95% CI
	5%	10%	20%	5%	10%	20%			
<u>Total test score criterion</u>									
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]
<u>Number of runs criterion</u>									
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 unidirectional 5%, 10%, and 20% cut off points corresponding to the innocent samples.

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

FIGURE ALSO PROVIDED AS .FIG FILE. THIS PAGE SERVES ONLY FOR REVIEW PURPOSES

Figure 1

Figure heading: Receiver Operating Characteristic (ROC) for correct total and alternations criterion for naïve and coaching condition.

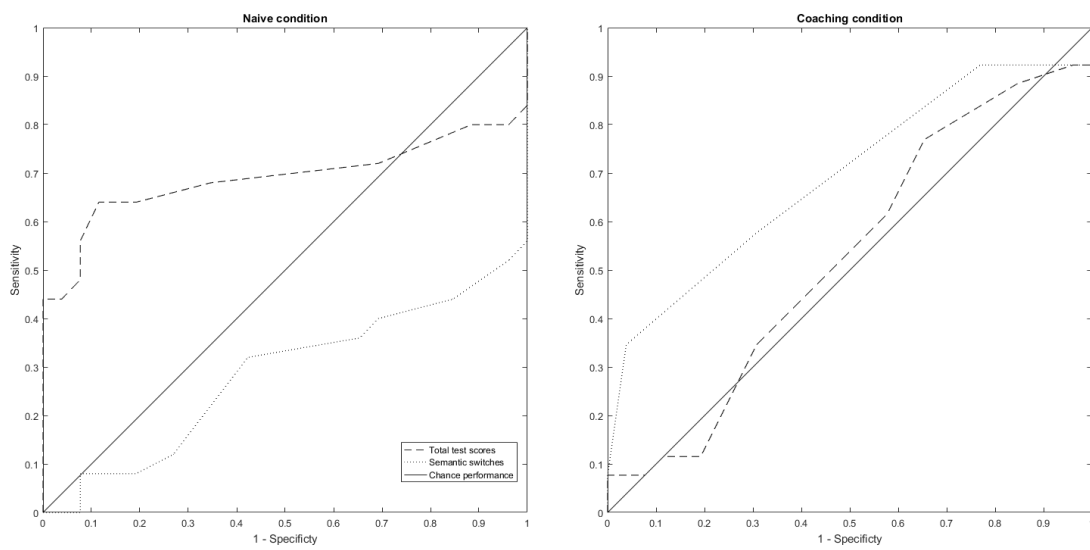


Figure notes: Note that ROCs in the naïve condition were aberrant. This is likely a consequence of the abnormal distribution of strategy levels used in this condition. In the coaching condition all participants reported using the same strategy level.