Resistance to coaching in forced-choice testing

Received: 19 December 2017 | Revised: 23 April 2018 | Accepted: 29 June 2018

DOI: 10.1002/acp.3443

RESEARCH ARTICLE

WILEY

Resistance to coaching in forced-choice testing

Robin Orthey[1,2] | Aldert Vrij[1] | Ewout Meijer[2] | Sharon Leal[1] | Hartmut Blank[1]

[1] Faculty of Science, University of Portsmouth, Portsmouth, UK

[2] Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

Correspondence
Robin Orthey, University of Portsmouth, Portsmouth, UK.
Email: robinorthey@googlemail.com

Summary

In forced-choice tests (FCTs), 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 nonrandom 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 and patterns 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 choose to randomize 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 randomize 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 alternation criterion and theoretical implications are discussed.

KEYWORDS

countermeasures, deception, forced choice

1 | INTRODUCTION

Forced-choice testing (FCT) has been used as a test to detect malingering of sensory impairment (Pankratz, Fausti, & Peed, 1975). More recently, its use has been extended to detect cases of faked memory loss (e.g., Denney, 1996; Hiscock & Hiscock, 1989; Pankratz, 1983; Van Oorsouw & Merckelbach, 2010) and concealed information (e.g., Giger, Merten, Merckelbach, & Oswald, 2010; Meijer, Smulders, Johnston, & Merckelbach, 2007; Orthey, Vrij, Leal, & Blank, 2017; Shaw, Vrij, Mann, Leal, & Hillman, 2012), from which guilty knowledge can be inferred. In the case of concealed information detection, a typical test works as follows: A suspect is presented with a series of questions about the crime. With each question, two equally plausible answer alternatives are presented: a correct and an incorrect one. For example, a question such as "What was the murder weapon" could be accompanied with two answer alternatives such as "gun" and "knife." Suspects are instructed to select the correct answer or guess if they do not 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, that is, 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 et al. (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 (Carmerer, Ho, & Chong, 2004). In cognitive hierarchy theory, 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, that is, 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 sequence 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 five 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 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, whereas the number of runs criterion will detect examinees who employ Level 2 strategies.

Specifically, in this study, we investigated two questions:

1. What is the effect of coaching on the strategies guilty and innocent participants select?
2. 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).

2 | METHOD

2.1 | Participants

One hundred four 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.

2.2 | Procedure

First, examinees were assigned to one of two virtual reality 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 three-dimensional 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 3 min 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 virtual reality 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 posttest 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%).

2.3 | 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 & Kirshenbaum, 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.

2.4 | Design and measures

This study featured a 2 (veracity: guilty vs. innocent) × 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 operating characteristic curve (ROC; Tanner & Swets, 1954), 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 and 1 with 0.5 indicating chance performance and a higher number meaning better discrimination between guilty and innocent examinees.

Participants' answers to the open questions about their behaviour during the test were categorized 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 randomizing between correct and incorrect answers were assigned to this level. Two blind and independent raters categorized the responses according to examples within each strategy level as specified in Orthey et al. (2017). Interrater reliability was high (89% absolute agreement). Responses that did not fit any category were omitted from the analysis (one participant).

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

3 | RESULTS

3.1 | Strategies

First, we examined the strategies examinees reported. We hypothesized that coaching would elicit higher level strategies in guilty examinees (Hypothesis 1). Table 1 depicts the frequencies of selected strategies divided by conditions. Innocent examinees reported using all types of strategies naturally, but when coached, they seemed to endorse either answering honestly or randomizing. 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.

A chi-squared test was performed, and we found a relationship between coaching and the used strategy level for guilty examinees, χ^2 (2, N = 51) = 16.32, p < 0.001. 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.

TABLE 1 Frequencies of strategy levels per condition

	Truth tellers		Liars	
Strategy level	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

Additionally, we analysed the detection accuracy of the correct total criterion per strategy level. Ninety per cent 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.

3.2 | Detection accuracy

We assessed detection accuracy for specific cut-off as well as the entire range of the 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 = 0.69, p = 0.020, 95% CI [0.53, 0.86]. In the coaching condition, the correct total did not distinguish guilty from innocent examinees better than chance, AUC = 0.53, p = 0.742, 95% CI [0.37, 0.69]. 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%, whereas specificity remained the same at 92.3%. At the 20% cut-off, sensitivity was 64% with a specificity of 88.5%. When coached, the sensitivity dropped to 7.7% with a 100% specificity at the 5% cut-off. At the 10% cut-off, sensitivity remained at 7.7%, but specificity declined to 92.3%. At the 20% cut-off, sensitivity was 11.5% with a specificity of 88.5%. This suggested a sharp decline in detection accuracy for the correct total criterion in case of coaching, which supports Hypothesis 2.

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 = 0.26, p = 0.008, 95% CI [0.14, 0.43]. However, in the coaching condition, the number of runs differentiated guilty from innocent examinees significantly better than chance performance, AUC = 0.69, p = 0.018, 95% CI [0.55, 0.84]. 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.

[1]Caution is warranted when interpreting these AUCs. The empirical ROCs are skewed (see Figure 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 overperformance not underperformance.

TABLE 2 Detection accuracy for the alternation criterion

Condition	Sensitivity 5%	Sensitivity 10%	Sensitivity 20%	Specificity 5%	Specificity 10%	Specificity 20%	AUC	*p*	95% CI
Total test score criterion									
Naïve	48%	56%	64%	92.3%	92.3%	88.5%	0.69	0.020	[0.53, 0.86]
Coached	7.7%	7.7%	11.5%	100%	92.3%	88.5%	0.53	0.742	[0.37, 0.69]
Number of runs criterion									
Naïve	0%	8%	8%	92.31%	92.31%	80.71%	0.26	0.008	[0.14, 0.43]
Coached	7.69%	34.62%	57.69%	100%	96.15%	69.23%	0.69	0.018	[0.55, 0.84]

Note. Sensitivity and specificity for number of runs criterion were based on the unidirectional 5%, 10%, and 20% cut-off points corresponding to the innocent samples.

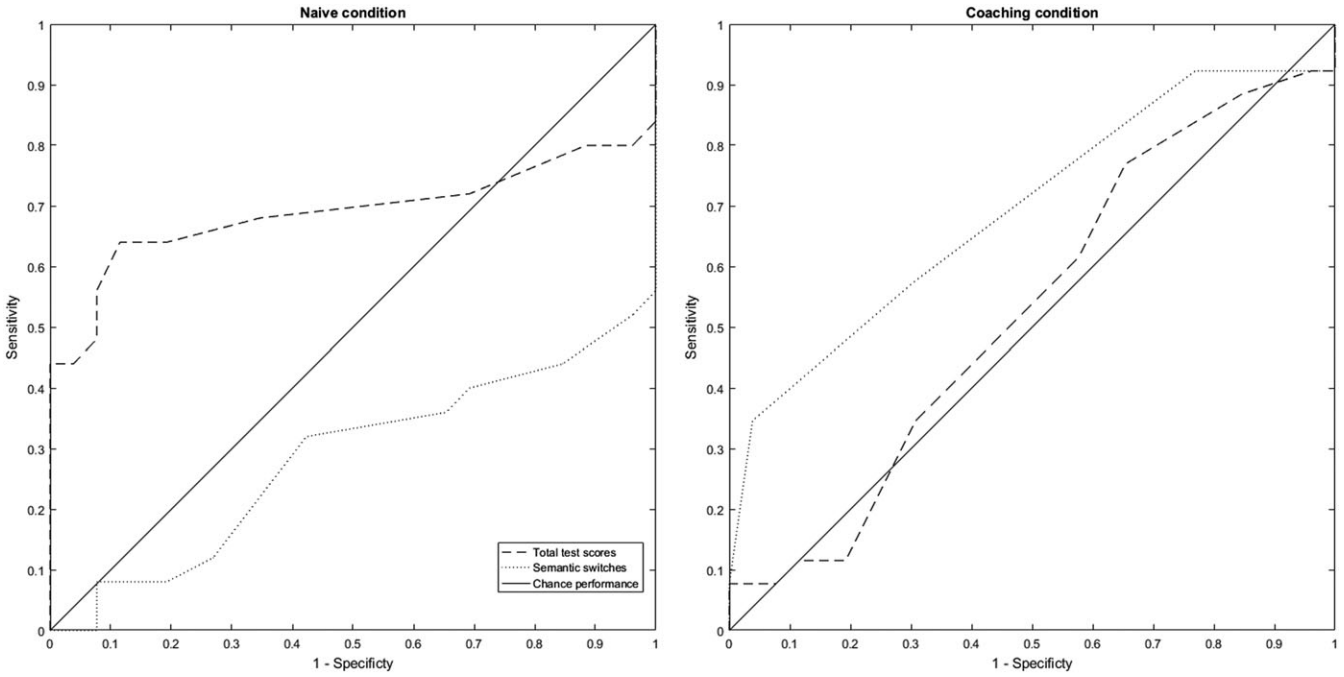

FIGURE 1 Receiver operating characteristic curve for correct total and alternation criteria for naïve and coaching conditions. Receiver operating characteristic curves in the naïve condition were aberrant. This is likely a consequence of the abnormal distribution of strategy levels used in this condition. 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 ($M = -0.05$, $SD = 1.18$) favoured alternating between correct and incorrect answer alternatives, but coached innocent examinees prioritized alternations between horizontal positions ($M = -0.46$, $SD = 0.91$).

3.3 | Incremental validity

Finally, we assessed the incremental validity of a two-step classification process. As Step 1, we used the correct total criterion with the conventional unidirectional cut-off at 5%. That is, all participants whose correct total score fell within underperformance were classified as guilty. As Step 2, 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

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

Condition	Sensitivity 5%	Sensitivity 10%	Sensitivity 20%	Specificity 5%	Specificity 10%	Specificity 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

Note. 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.

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.

4 | 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 behaviour (see Zvi, Nachson, & Elaad, 2012, and Zvi, Nachson, & Elaad, 2015, for similar findings in 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% and 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 an 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 randomize between correct and incorrect answer alternatives, whereas innocent examinees were more inclined to randomize 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 overfitted. 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 alternation 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 randomization 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. Second, 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 (Ericsson & Simon, 1980; Nisbett & Wilson, 1977; Schwarz, 1999). The concern is that human subjects may not be aware of the true reasons of their behaviour 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 behaviour (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 behaviour 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.

ORCID

Robin Orthey http://orcid.org/0000-0002-6185-3061
Aldert Vrij http://orcid.org/0000-0001-8647-7763

REFERENCES

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. https://doi.org/10.1080/13854046.2014.978383

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

Denney, R. L. (1996). Symptom validity testing of remote memory in a criminal forensic setting. *Archives of Clinical Neuropsychology*, *11*(7), 589–603. https://doi.org/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.

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. https://doi.org/10.1080/15228932.2010.489875

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.

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. https://doi.org/10.1016/j.acn.2003.07.004

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. https://doi.org/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 and Law*, *8*, 311–318. https://doi.org/10.1080/1068316021000054256

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, *109*(2), 330–357. https://doi.org/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. https://doi.org/10.1002/acp.3310

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. https://doi.org/10.1037/h0076722

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.htm

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*, *19*, 240–254. https://doi.org/10.1111/j.2044-8333.2012.02070.x

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

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401–409. https://doi.org/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*, 1–13. https://doi.org/10.1027/1015-5759/a000438

Van Oorsouw, K., & Merckelbach, H. (2010). Detecting malingered memory problems in the civil and criminal arena. *Legal and Criminological Psychology*, *15*, 97–114. https://doi.org/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 and Law*, *14*(6), 523–528. https://doi.org/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.

How to cite this article: Orthey R, Vrij A, Meijer E, Leal S, Blank H. Resistance to coaching in forced-choice testing. *Appl Cognit Psychol*. 2018;32:693–700. https://doi.org/10.1002/acp.3443