UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Orthey, R., Meijer, E., Kooistra, E., & Broers, N. (2022). How to Detect Concealed Crime Knowledge in Situations With Little Information Using the Forced Choice Test. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.37483

Cognitive Psychology

How to Detect Concealed Crime Knowledge in Situations With Little Information Using the Forced Choice Test

Robin Orthey[1] [a], Ewout Meijer[2], Emmeke Kooistra[3], Nick Broers[2]

[1] Aoyama Gakuin University, Tokyo, Japan, [2] Maastricht University, Maastricht, Netherlands, [3] Nederlands Studiecentrum Criminaliteit en Rechtshandhaving, Amsterdam, Netherlands

Keywords: Forced Choice Test, Concealed Information Test, Crime Amnesia

https://doi.org/10.1525/collabra.37483

Collabra: Psychology

Vol. 8, Issue 1, 2022

The Forced Choice Test (FCT) can be used to detect concealed crime knowledge, but it requires more evidence than typically available from a crime to be constructed. We propose a method to repeat individual pieces of evidence to achieve the necessary test length, hence widening the practical applicability. According to our method, FCT trials are created so that on each trial examinees are presented with a novel and unique decision between two answer alternatives even if a specific piece of information is presented again. We argue that if the decision in each trial is unique, the properties and diagnosticity of a traditional FCT can be maintained. In experiment 1, we provide a proof of concept by comparing our novel method with a traditional FCT and demonstrate that an FCT with repeated presentation of the same evidence has diagnostic value (AUC = .69) albeit less so than a traditional FCT (AUC = .86). In experiment 2, we put our novel FCT to the test in a situation with insufficient information for a traditional FCT alongside the Concealed Information Test (CIT), which also detects concealed information but relies on psychophysiological indices. Both, the FCT (AUC = .81) and CIT (AUC = .83) were diagnostic and combining them increased the detection accuracy even further (AUC = .91). If replicated, our novel FCT increase practical applicability of the FCT in general and in conjunction with the CIT.

1. Introduction

The idea behind concealed knowledge detection tests is to identify a suspect as the perpetrator of a crime by demonstrating that the suspect is in possession of knowledge only the perpetrator can have. The Forced Choice Test (FCT) can be used for this purpose (Bianchini et al., 2001; Binder et al., 2014; Denney, 1996)

and has a straightforward procedure. It begins by creating questions about facts from the crime and for each question two possible answer alternatives are generated. One alternative is always correct and is derived from the evidence of the investigation, while the other is always incorrect. During the test procedure, the examinee faces the questions, corresponding answer alternatives, and a simple set of instructions. Their task is to select the answer they believe to be correct for each question or guess in case they don't know. The FCT distinguishes examinees with- and without knowledge of the crime through their response strategy during the test. Examinees without concealed knowledge can only guess, meaning they have a 50% chance to select the correct answer for each question, leading to a total test score that resembles chance performance. In contrast, examinees with knowledge of the crime are expected to recognize the correct answer and tend to select incorrect answers on purpose, leading to below chance level performance (e.g., Merckelbach et al., 2002). This behaviour is known as underperformance and is used as criterion for concealed knowledge.

One benefit of the FCT is that performance of an individual without crime knowledge follows a known – binomial – distribution. As a consequence, the decision threshold for classifications in individual cases can be set to a specific predefined false positive rate. In practice, a 5% false positive rate is typically deemed acceptable (e.g., Denney, 1996; Pankratz et al., 1975), and has been used by many researchers. For example, Merckelbach et al. (2002) constructed an FCT with 15 trials and detected 40% of participants who feigned autobiographical memory loss. Using the same procedure, Verschuere, Meijer, & Crombez (2008) detected 58% of their sample with an FCT that had 25 trials. Others asked participants to feign memory loss for a mock crime. Shaw, Vrij, Mann, and Hillman (2014) used an FCT with 12 trials and detected 42% of their 86 participants who committed the mock crime. Similarly, Jelicic, Merckelbach, and van Bergen (2004) detected 59% with a 25 trial FCT, and Giger, Merten, Merckelbach, and Oswald (2010) report

a Corresponding author: robinorthey@googlemail.com

a 45% detection rate for an FCT with 19 trials. Meijer, Smulders, Johnston, and Merckelbach (2007) report two FCT studies. In study 1, university students commit a mock crime, and a 12 trial FCT could detect concealed knowledge in 47% of the sample. In study 2, the authors detected 63% of a community sample for autobiographical details with a 12 trial FCT. Thus, the FCT has been found to detect substantial proportions of examinees with concealed knowledge at low pre-defined false positive rates.

But there is a catch. A minimum amount of information is needed to create an FCT. Some authors recommend a minimum of 25 questions (Denney, 1996), while others argue 12 questions are sufficient (Van Oorsouw & Merckelbach, 2010). Test length in real life is, however, dictated by the information available from the investigation which may fall well short of the desired 25 or even 12. Additionally, any information that has become public knowledge cannot be used as the suspect may have gained this knowledge via other means such as media outlets (Podlesney, 2003). In general, longer test sizes are desirable due to the improved reliability (Hambleton & Cook, 1983), and increases in test length have been associated with better diagnostic accuracy (Lieblich et al., 1974; Lukács, 2021) in the Concealed Information Test (CIT; Lykken, 1959, 1960, 1974), another test that detects concealed knowledge using physiological indices or response times. It is also noteworthy that FCTs with more trials tend to feature higher detection accuracy (e.g. Jelicic et al., 2004). The only exception is study 2 from Meijer et al. (2007), who achieved a high detection accuracy with a test length of 12 items. However, unlike other FCT studies, it was conducted with a community sample rather than university students, which may be another reason for the high detection accuracy because higher education could be associated with a greater likelihood to see through the mechanism of the FCT.

Furthermore, there are two specific reasons why a short test length is problematic for the FCT. First, short FCTs are less capable of detecting underperformance, because the most extreme test results, that is, zero correct items selected, feature probabilities according to chance that fall short of the commonly used 5% or more conservative 1% (Orthey, 2019) criterion. Second, additional criteria that can be added to the FCT such as the runs test – a test that reflects the probability of alternations between correct and incorrect answers - This test draws on the human inability to replicate true randomness (Nickerson, 2002; Wagenaar, 1972) and Verschuere et al. (2008) suggest using the runs test to detect intentional random responding, a popular alternative response strategy in the FCT for examinees with concealed knowledge. However, in their study the runs test was ineffective, but Orthey, Vrij, Meijer, Leal, & Blank (2018) argue that the FCT used by Verschuere et al. (2008) did not have the necessary test length to reliably discriminate real from replicated randomness. Hence, test length is a crucial determinant for the effectiveness of the FCT to detect concealed knowledge.

In this manuscript we try to increase the applicability of the FCT by exploring a novel way to construct answer alternatives pairs. When insufficient evidence is available to construct a traditional FCT, we propose to reuse pieces of evidence to increase the potential test length. Simply repeating a traditional FCT pair of alternatives is unlikely to work. When faced with an FCT about a specific event, the examinee must deduce the correct answer to execute their response strategy. For examinees with concealed knowledge this is simply a memory recognition process. But for those without concealed knowledge who are motivated to select the answers they believe to be correct (Orthey et al., 2017), we can expect that the response for a repeated pair will be the same. In other words, the response for a repeated pair is not independent from the previous presentation and will therefore not add information. We hypothesized that evidence can be reused as long as examinees are never faced with the same choice, so if an answer alternative is presented in a different context or question examinees may not be consistent with their prior choice, but still perform to the best of their ability. To this end we modified the traditional FCT procedure in two ways. First, we replace traditional FCT questions such as "What weapon was the offender carrying?" or "Which item did the offender steal?" with the more generic question "Which answer is more related to the crime?". Correspondingly, traditional pairs from the same category, for example, "gun" and "knife" or "wallet" and "watch", are replaced with two seemingly unrelated answers, for example, "knife" and "watch". This way, more unique trials can be derived from the same number of pieces of evidence, even though the proportion each unique piece of information takes up in the FCT remains unchanged. For example, the following three answer alternative pairs could be used in a traditional FCT: "Gun" & "Knife"; "Shirt" & "Pants"; and "Wallet" & "Watch". Subjected to our method these three pairs generate the following six unique combinations: "Gun" & "Pants"; "Gun" & "Watch"; "Shirt" & "Knife"; "Shirt" & "Watch"; "Wallet" & "Knife"; "Wallet" & "Pants". This way, we maintain the underlying FCT structure, namely two equally plausible answers of which one is related to the crime and the other is not, while increasing the test length by repeated presentation of the evidence. We expect examinees without concealed knowledge not to default to prior choices when they encounter a piece of evidence again, because the task is framed as relational, "Which of these answers is more related to the crime?", rather than factual and the presented pair of answers is always unique. We investigate here if and how effective this method is.

In study 1 we provide a proof of concept for this this novel FCT type by comparing the diagnostic accuracy of our novel answer alternative pairs with that of a traditional FCT. Then in study 2, we validate our findings in a more ecologically valid situation, and compare our novel FCT with a psychophysiological knowledge detection test, the Concealed Information Test.

2. Method

2.1. Participants

A total of 160 undergraduate students (*mean* age = 21.88, *SD* = 3.51) participated in this experiment (107 male; 53 female). Participants received either course credit or €5 as compensation. Ethical approval was obtained.

2.2. Forced Choice Test

The traditional FCT featured 20 pairs of categorically related pictorial answer alternatives, for example, a picture of a handgun paired with a picture of a rifle, or two pictures of public places. The pairs were checked for equal plausibility in Orthey et al. (2018). The FCT with repeated alternatives used the same pictorial answer alternatives but paired the correct answer from each related pair with an incorrect answer from a different pair. As a result, a total of N * (N – 1) unique pairs can be generated, where N denotes the number of unique pieces of information available to the examiner.

Thus, 20 unique pairs can yield a total of 380 unique combinations of a correct and incorrect answer alternative. However, to retain a reasonable duration of the experiment we limited our test length to 100 pairs, with each of the 20 correct and incorrect answer alternative repeated 5 times in a counterbalanced fashion.

The procedure in our experiment for the standard and novel FCT deviated from the traditional FCT procedure in two ways. First, instead of specific questions referring to a single category, for example, "What kind of weapon was found in the apartment?", all answer alternative pairs were preceded by the same generic question, "Which of these two pictures is more related to the terrorist's apartment?". Second, a salient webcam was placed on top of the computer screen. During the set-up phase the experimenter would go through great length to adjust the camera position to point at the participants' face and participants were instructed to keep their attention on the computer screen. No actual video footage was recorded, and the real purpose of this procedure was to induce the belief that facial expressions were relevant to the test as there is a common (incorrect) belief that these are diagnostic cues to deception (e.g., Global Deception Research Team, 2006). By shifting the participants attention to their facial expressions, the webcam served as a misdirection to mask the underlying mechanism of the FCT. We did this because Shaw et al. (2014) demonstrate that participants who understand the FCT's rationale are more successful at avoiding detection.

2.3. Design and Measures

This study featured a 2 (Knowledge: knowledgeable vs unknowledgeable) x 2 (Pair type: traditional vs repeated) between subject design with the number of correct answers selected as dependent variable. We computed the FCT score twice for the repeated FCT. Once, we used only the first 20 trials, so that both FCTs were of equal length. Then, we also computed the FCT score over all 100 trials. The FCT total scores were transformed into z-values according to Siegel's (1956) formula for binomial distributions, with negative scores indicating more incorrect answers than expected by chance, while positive scores suggested more correct answers than expected by chance. Orthey, Vrij, Leal, and Blank (2017) demonstrated that some participants with concealed knowledge tend to endorse correct answers rather than avoid them and that their prevalence is increased in the presence of misdirection manipulations. Therefore, we transformed the z-scores into absolute values, meaning the larger the score the less likely the total score was to occur through chance.

Diagnostic accuracy was expressed using Signal Detection Parameters (see Hanley & McNeil, 1982). Sensitivity refers to the detection rate of examinees with knowledge of the mock crime, while specificity refers to the detection rate of examinees without this knowledge. Sensitivity and specificity require a cut-off point, and we used the traditional cut-off corresponding to a 95% specificity, defined at z-scores of ⩾ 1.96, and a more conservative cut-off corresponding to 99% specificity, defined at z-scores of ⩾ 2.58, as suggested by Orthey (2019). Bidirectional cut off points were used, because our dependent variable reflected absolute test scores. We also computed the Receiver Operating Curve (ROC), which plots sensitivity against specificity for all possible cut offs. The Area Under the Curve (AUC) of the ROC can be used as a general detection accuracy indicator. It ranges between 0 and 1 with 0.5 representing chance performance. The higher the value, the better is the discriminant ability of the test (Hanley & McNeil, 1982; Tanner & Swets, 1954). We compare the AUCs between conditions using the bootstrap method for unpaired AUCs with the pROC R module (Robin et al., 2011).

To check the response strategies of knowledgeable participants we recoded their answers on the questions "How did you decide on each trial what answer to select?" into the strategy levels 0 – 2 suggested in Orthey et al. (2018). Put simply, level 0 represents endorsement of correct answers, level 1 represents avoidance of correct answers, and level 2 represents randomisation between correct and incorrect answers. Response strategies were coded by two independent raters for all participants. In cases of disagreement both raters had the opportunity to discuss the case and then decided for themselves if they made a change or not. Inter-rater reliability post discussion was high (absolute agreement: 91.87%; Cohen's Kappa = .89). We used the codings of the rater who collected the data.

2.4. Procedure

First, participants were assigned to one of two Virtual Reality (VR) scenarios from Orthey et al. (2018) in a counterbalanced fashion that ran in the Unity 3D editor (https://unity.com/). In one scenario participants had to take the role of an intelligence officer and investigate the apartment of a terrorist with the objective to find as many clues about an imminent attack as possible. Details (for example, a weapon hidden inside a closet) were highlighted with a bright yellow exclamation mark, and the simulation terminated when all clues were found. After the simulation participants received the following instruction to hide their knowledge from the mock crime: "In the subsequent part, pretend that you do not know anything about the terrorist's apartment". Details from this scenario were treated as crime relevant knowledge. In the other scenario participants took the role of a real estate agent and were asked to evaluate a different apartment. The simulation terminated after 3 minutes, and no other instructions were provided. Participants in this group were unaware of all clues present in the other condition.

Table 1. Frequencies of strategy levels for participants with concealed knowledge and mean and standard deviation of FCT z-scores per strategy level per answer alternative type

	Traditional			Repeated		
	Level 0	Level 1	Level 2	Level 0	Level 1	Level 2
Frequency / proportion	11 / 27.5%	7 / 17.5%	22 / 55%	11 / 28.21%	7 / 17.94%	21 / 53.8%
Mean *z*-score	3.15	-2.01	0.12	5.59	-4.3	1.54
SD *z*-score	2.23	2.66	2.17	3.67	2.38	3.21

Notes. Level 0 predicts endorsement of correct answers, level 1 predicts avoidance of correct answers and level 2 predicts randomisation between correct and incorrect answers.

Next, all participants were subjected to the FCT. Participants were instructed that they would be presented with questions about the imminent attack and that each question came with two answer alternatives. The instruction was to select the correct answer or guess if they did not know it. Participants were unaware of the total number of questions and were assigned to either a traditional FCT or an FCT with repeated answer alternatives (see below for details) in a counterbalanced fashion, participants were equally distributed over all condition combinations, resulting in 40 participants per condition.

After the FCT participants were instructed to answer all subsequent questions honestly. Participants were questioned about their strategies to appear innocent on the FCT with the question: '*How did you decide on each trial what answer alternative to select?*' Their verbal response was audio-recorded and later recoded as described in the previous section.

Finally, a memory test about the details from the mock crime scenario was administered to participants in the concealed knowledge condition to check if they had the knowledge they were instructed to hide. All participants received the questions and two answer alternatives of the traditional FCT again with the explicit instruction to answer honestly. The option 'I don't know' was added as a third option to all answer alternative pairs. Memory performance was satisfactory (*mean* = 91.5% correct) and no participants were excluded due to poor performance on this part.

3. Results

3.1. Strategies

Table 1 displays the frequency of response strategies, associated mean z-scores and their standard deviation for per FCT type for participants with concealed knowledge. Randomisation between correct and incorrect answers (level 2) were the most prevalent strategies representing more than half of the sample in both conditions. Additionally, in both conditions, endorsement of correct answers (level 0) was the next most prevalent strategy with close to 30% of the sample reporting it, followed by avoidance of correct answers (level 1) with slightly less than 20% of the sample. As expected, test scores of participants who reported endorsement of correct answers were high, while those who reported avoid correct answers had low scores. Participants who reported randomisation between correct and incorrect answers scored around chance level.

3.2. Detection accuracy

The ROC curves are presented in Figure 1. The AUC scores showed the FCT could distinguish knowledgeable from unknowledgeable participants better than chance in both conditions. The traditional FCT (*AUC* = .86, $p < .001$, *95% CI* = [.78 - .94]) had the best detection accuracy and the FCT with repeated answer alternatives had a lower yet above chance detection accuracy. (*AUC* = .69, $p = .002$, *95% CI* = [.58 - .80]). The AUC of the traditional FCT was not significantly larger than that of the repeated FCT with 100 trials, $D = 1.87$, $df = 143.56$, $p = .063$, but it was significantly larger than that of the repeated FCT with the first 20 trials, $D = 2.36$, $df = 146.46$, $p = .020$. The detection accuracy estimates corresponding to an expected specificity of 95% and 99% can be found in Table 2. Together, the data supports our hypothesis than an FCT with repeated answer alternatives can detect concealed knowledge better than chance.

4. Discussion

In our efforts to develop a new FCT procedure for situations that feature an insufficient amount of information for a traditional FCT, we first checked if information can be reused in an FCT design. We found that both a traditional FCT and an FCT that repeatedly presented the critical information detected concealed knowledge better than chance, providing a proof of concept.

Diagnostic accuracy of our novel FCT procedure was inferior to a traditional FCT. This difference should, however, be seen in the light of the superior performance of the traditional FCT rather than the inferior performance of our novel FCT procedure. In line with previous experiments such as Merckelbach et al. (2002), Giger, et al. (2010), study 1 and 2 of Meijer et al. (2007), Shaw et al. (2014), and Jelicic et al. (2004) both types of FCT were able to detect meaningful proportions of participants with concealed knowledge at high specificity levels. One of the potential explanations for the relatively high accuracy in the traditional FCT is our use of the misdirection manipulation paired with counting both under- and over performance as diagnostic of having concealed information or the switch from crime specific questions to a single generic one.

A possible explanation for the difference between the traditional and repeated FCT using the first 20 trials could be that the former contained more crime related information due to the random sequence of trials. Furthermore, the amount of crime information presented in the first 20 trials may have differed among participants due to the ran-

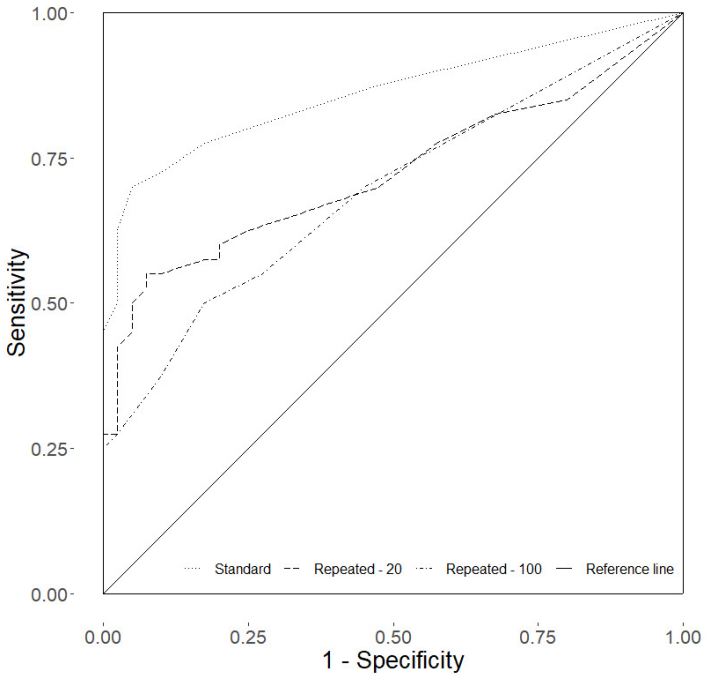


Figure 1. Receiver Operator Characteristic for the traditional and repeated FCT.

Table 2. General and single cut off detection accuracy per FCT answer alternative type.

	1% cut-off		5% cut-off				
	Sensitivity	Specificity	Sensitivity	Specificity	*AUC*	*p*	95% *CI*
Traditional	50%	97.5%	70%	95%	.86	< .001	[.78-.94]
Repeated - 20	25%	100%	37.5%	90%	.69	.002	[.58-.80]
Repeated - 100	52.5%	92.5%	57.5%	82.5%	.72	<.001	[.60-.84]

Notes. 1% and 5% correspond to z-scores ⩾ 1.96, ⩾ 2.58, respectively. Repeated 20 refers to the repeated FCT using only the first 20 trials that were presented. Repeated 100 refers to the repeated FCT using all trials.

dom presentation of trials. However, when looking at all 100 trials that contain all 20 pieces of information, we see a considerable increase in sensitivity with a disproportionally small decrease in specificity, and both repeated FCTs feature a similar general diagnostic accuracy above chance level. Still, the traditional FCT featured the best detection accuracy on all accounts, which may suggest that the ratio of information to test length may also affect detection accuracy.

In sum, we provide a proof of concept that it is possible to create a diagnostic FCT with repeated information from the crime. However, this study was conducted in a laboratory setting and does not aptly reflect a situation with a lack of evidence. Therefore, we conduct a follow-up study that features insufficient information for a traditional FCT to better assess the value of a repeated FCT in such situations.

5. Study 2

A more ecologically valid situation would feature so few pieces of information that a traditional FCT cannot be constructed. Therefore, rather than comparing our novel answer alternative pairs to a traditional FCT, we use the Concealed Information Test (CIT) as a reference. The CIT also detects concealed crime knowledge and uses a multiple-choice format with, typically, five answer alternatives. In contrast to the FCT, concealed knowledge is not inferred from a selection pattern, but from physiological responses to the answer alternatives. The underlying idea of the CIT is that knowledgeable examinees will display an elevated physiological response when presented with the correct answer to the question (Lykken, 1959, 1960, 1974). This differential response does not occur in unknowledgeable examinees as all answers appear as possible answers. Meta-analyses demonstrate that the diagnostic accuracy of the CIT is good (Ben-Shakhar & Elaad, 2003; Meijer et al., 2014). Evidence further suggests that combining the FCT and CIT has a better diagnostic accuracy than either test alone. In two studies, Meijer et al. (2007) also combined the CIT and FCT scores. In study 1, this led to a 100% correct classification rate for participants without concealed knowledge and a 57% detection rate for participants with concealed knowledge. In study 2, the combination of CIT and FCT scores produced an 88% detection rate for participants with concealed knowledge. In both studies the combined score exceeded the individual CIT and FCT scores.

The FCT and CIT both detect concealed knowledge and can be constructed from the same material, with the caveat that more evidence is required for the FCT than the CIT. Meijer, Verschuere, and Ben-Shakhar (2011) for example, recommend at least three to five CIT questions. While it is impossible to construct a traditional FCT with that amount of information, our repeated FCT should be able to achieve a satisfactory test length. Thus, to estimate the diagnostic validity of the repeated FCT in a situation with not enough evidence for a traditional FCT, we subjected our participants to a CIT and repeated FCT examination about a mock terror organization. We expect that both, the CIT and our novel FCT, can detect concealed knowledge better than chance.

6. Method

6.1. Participants

Eighty-one undergraduate students were recruited for this experiment. One participant was excluded due to poor memory performance (see below). Eighty participants remained in the sample; their *mean* age was 23.53 years (*SD* = 5.63) and 17 were men. Participants received 10€ or course credit as compensation. The monetary reward was advertised as a reward of 7.50€, with the opportunity to win an additional 2.50€. Regardless of their performance all participants received the additional 2.50€ as compensation.

6.2. Procedure

All participants received a file about a mock terror organisation. This file contained information such as names, places, plans of attack and other details. Each detail was embedded in a single paragraph of two or three sentences. In addition, some details were also presented in pictorial form. For example, the name of the terror organisation was also displayed on a picture of their fictional flag. Some parts of the file were intentionally blacked out to imply that this information was redacted. The files differed in the degree of redaction per condition. For half of the participants (N = 40), all relevant details were redacted, while for the other half (N = 40) all but five pieces of information were redacted. Participants in the former condition are referred to as unknowledgeable and participants in the latter as knowledgeable. Then, all participants were instructed to deny any knowledge gained from the file for the following procedure. As motivational incentive, participants were told they could receive additional monetary reward of 2.50€ if they were successful in doing so. However, all participants received the additional reward regardless of their test performance, including participants who received course credit.

The first task consisted of a test aimed to extract a hidden location mentioned in the mock crime file that was unrelated to the research question in this manuscript. The results of this test will be reported elsewhere. After that, a CIT examination was conducted followed by the FCT (both detailed below). The CIT was always conducted first to avoid any confounds as a result of participants' choices in the FCT. That is, having selected an answer in the FCT may result in increased significance and a biased CIT. Finally, participants received a questionnaire about their strategies to avoid detection during the CIT and FCT and completed a memory test to determine whether they remembered the five relevant pieces of information from the fake terrorist file. Memory performance was good, on average participants remembered 4.7 (*SD* = 0.46) from 5 pieces of information. We excluded knowledgeable participants who correctly remembered less than four of the five items from the mock crime and one participant was excluded for this reason. The experiment was conducted by two experimenters. One, who would supply the file and instructions to feign ignorance, and the other, who conducted the FCT and CIT examination blind to the participants' condition. The first experimenter also supplied the post test questions and memory check.

6.3. CIT & FCT

The CIT examination followed the standard protocol in the literature (e.g., Matsuda et al., 2019) as it resembles best the current application of the CIT by the police in Japan. We formulated five questions about critical details from the terrorist file, for example "What is the estimated number of members of this terror organisation?". For each question we generated five additional incorrect answer alternatives. With the correct answer this resulted in six alternatives per question. To avoid biases within our sets of answer alternatives we had to make sure all answers were equally plausible (see Doob, & Kirschenbaum, 1973). To do so we submitted the CIT questions and answers to a pilot procedure. Over two rounds (N1 = 30; N2 = 20) we asked people unfamiliar with the mock crime information to choose the answer they believed to be correct for each question. If on any question an answer was selected by more than 40% of respondents the set of five answers was deemed biased (similar to Merckelbach et al., 2002) and a new set of answer alternatives was generated. At the end of the second cycle all sets of answer alternatives passed this criterion. Before the start of the test, we showed participants all questions with answer alternatives and described the test procedure (Verschuere & Crombez, 2008). In practice, this step is used to exclude questions that are affected by leakage of information, for example through news or trial proceedings (e.g., Denney, 1996). The procedure was as follows:

First, each question was presented on the screen for ten seconds. Then, the six answer alternatives were presented in a random order with one exception, namely that for each question, one of the incorrect answers was always presented first. This answer was used to absorb the startle response, which is an elevated physiological response generated by the start of the sequential presentation (Gati & Ben-Shakhar, 1990; Meijer et al., 2011). This answer was excluded from the analysis. Answer alternatives were presented for six seconds and participants had to audibly respond with 'No' to each alternative. A 10 – 15s inter stimulus interval was added between the answer alternatives.

Participants' Skin Conductance Response (SCR) and Respiration Line Length (RLL) were recorded during the CIT. SCR was measured using a 16-bit DC 0,5-V system. Two Beckmann 0,8 cm Ag/AgCL electrodes were placed on the medial phalange of the index and middle finger of the non-dominant hand. The electrodes were filled with isotonic electrode paste (0.9% NaCl). RLL was measured using a respiratory band positioned on the bottom part of the sternum.

All data were acquired using BrainVision bio-amplifiers with a sample rate of 1000 Hz (Version 2.1, BrainVision LLC, Morrisville, NC, www.brainvision.com).

For the FCT participants received the following instructions: "*Next, you will be presented with pairs of words. One word is always more related to the terror organisation than the other. On each trial select the word most related to the terror organisation. Use the mouse to select the answer*". We created two types of trials using the answer alternatives generated for the CIT. First, the critical trials contained one of the correct answers and one incorrect answer. Critical trials were generated as in study 1. Every correct answer alternative was paired with the incorrect answer alternatives from the other questions, resulting in a total 20 critical trials. The other type of trials were filler trials (e.g., Jelicic et al., 2004). Filler trials did not contain a correct answer and their purpose was to mask the FCT test mechanism. We created 42 filler trials using the remaining incorrect answer alternatives from the CIT answer alternative sets and two other potential CIT questions that were discarded during the pilot procedure. From each discarded question we selected two equally probably choices. Filler trials were constructed in the same manner as the critical trials. In total the FCT featured 62 trials, 20 critical trials and 42 filler trials. Both types of trials were presented interchangeably in random sequence and the position of the correct answer (left/right side) was determined randomly on each trial. The SCR and RLL sensors remained attached to the participant and just as in study 1 participants were made aware of a camera directed at their face during this procedure to further distract from the FCTs test mechanism.

6.4. Design

This experiment featured a single between-subjects factor Information (knowledgeable vs unknowledgeable) with the FCT and CIT measures as dependent variables. For the FCT, we summed the number of correct items selected in the critical trials and then z-transformed them according to the binomial distribution. Filler trials were omitted from the analysis as they did not feature correct answers. The standardized scores were then transformed into absolute scores for the same reasons as outlined in study 1. Hence, the higher the FCT score, the less likely it was to occur through chance.

We computed three different CIT indicators: the SCR, RLL, and SCR and RLL combined score. SCR and RLL values were computed for each stimulus that was presented. SCR was determined as the difference between the minimum and maximum SCR in the time interval from 1 to 5 seconds following stimulus presentation. If the SCR only declined in that time window it was set to zero. For the RLL we followed a specific procedure to minimize bias by the position of the respiration measure at epoch onset (see Elaad et al., 1992; Matsuda & Ogawa, 2011). We extracted ten 13-second epochs following stimulus presentation. The epochs varied in their starting position with an incremental 0.1 seconds per epoch to avoid bias of the phase of the respiration at stimulus presentation. We computed the length of the respiration line for each epoch and then averaged this value over the ten epochs, resulting in the RLL value per answer alternative. Next, we standardized the SCR and RLL value within each question (Ben-Shakhar, 1985). Specifically, we standardized the value of the correct response to the entire set of answer alternatives excluding the answer alternative that was always presented first to absorb the startle response. Finally, we averaged the standardized SCR and RLL response to the correct answer over all five questions, resulting in a single SCR and RLL value per participant. A participant was classified as a SCR non-responder when the standard deviation over all CIT items was smaller than .01 (e.g., Klein Selle et al., 2019). Non-responders were excluded for the SCR measures. In total six participants without concealed knowledge and two participants with concealed knowledge were excluded, leaving 34 and 38 participants respectively. The RLL scores were multiplied by -1, because we expect the correct answer to elicit a shorter RLL (Meijer et al., 2014). A combined CIT score was computed by simply summing the SCR and RLL scores (Ben-Shakhar & Elaad, 2002; Elaad et al., 1992). The scores for SCR non-responders were set to zero for this operation, hence effectively only representing the RLL score. Finally, we computed a score for the FCT and CIT together, by summing the absolute FCT scores with the combined CIT score.

Diagnostic accuracy was assessed as in study 1. The AUC served as indicator of general diagnostic accuracy for the CIT and FCT. We compared AUCs using the Delong, Delong, and Clarke-Pearson (1988) method for paired AUCs with the pROC module (Robin et al., 2011). We test if FCT and CIT differ from each other (bidirectional) and if the combination of both tests has incremental validity over either test alone (unidirectional). In addition, we determined the sensitivity and specificity for the FCT corresponding to an expected specificity of 99% and 95% ($z \geqslant 2.58$; $z \geqslant 1.96$).

Participants' response strategies for the FCT were measured with a multiple-choice question asking about what participants did with correct answers during the FCT. One option implied endorsement of correct information and was recoded to a level 0 response strategy, another option suggested avoidance of correct answers and was recoded to a level 1 response strategy and the last option suggesting mixture of correct and incorrect answers was recoded to a level 2 response strategy.

7. Results

7.1. Strategies

Table 3 depicts the frequency of the three strategy levels for participants with concealed knowledge for the FCT. Per strategy level it also indicates the mean z-scores and their standard deviation. Level 1 strategies, avoidance of correct information, was the most prevalent, followed by level 2 strategies, providing a mixture of correct and incorrect answers. Level 0 strategies, endorsing correct information, occurred only once. The mean z-values per strategy level were in line with the expected direction.

7.2. Detection Accuracy

Table 4 depicts the general detection accuracy of the FCT, CIT measures, and combination of FCT and CIT with the AUC. All five indicators detected concealed knowledge

Table 3. Frequency / Proportion of the self-reported three strategy levels, and *mean* and *Standard deviation (SD)* of FCT z-scores of participants with concealed knowledge per strategy level

	Level 0	Level 1	Level 2
Frequency / Proportion	1 / 2.5%	22 / 55%	17 / 42.5%
Mean	4.25	-3.44	-0.30
SD	-	1.02	1.30

Notes. Level 0 predicts endorsement of correct answers, level 1 predicts avoidance of correct answers and level 2 predicts randomisation between correct and incorrect answers.

Table 4. General detection accuracy for the FCT and CIT measures.

	AUC	*p*	*95% CI*
FCT	.81	< .001	[.73 - .94]
SCR	.85	< .001	[.76 - .93]
RLL	.67	.008	[.54 - .79]
SCR + RLL	.83	< .001	[.73 - .92]
FCT + CIT	.91	< .001	[.85 - .97]

Notes. FCT = Forced Choice Test; SCR = Skin Conductance Response; RLL = Respiration Line Length; SCR + RLL = sum of SCR and RLL scores.

better than chance (ps < .05). RLL yielded the lowest effect size, while both FCT and SCR featured a higher diagnostic accuracy. We also computed sensitivity and specificity for the FCT for a single cut-off. Under the assumption of an a priori 99% specificity cut-off we observed a sensitivity of 42.5% and a specificity of 100%, with the traditional 95% specificity cut-off sensitivity was 57.5% and specificity 85%. The ROC of the FCT (see Figure 2 left) indicated a non-normal distribution of test scores evidenced by the relative high sensitivity in the high specificity region. In contrast, the CIT measures appear to be normally distributed (see Figure 2 right). The data supports our hypothesis that an FCT reusing information has diagnostic validity with as little as five pieces of information. Finally, we explored the combined diagnostic accuracy of the FCT and CIT's best measure by summing the FCT z-score and mean SCR-score. Non-responders' contribution to this sum was set to zero. The combination of both FCT and SCR yielded the best diagnostic accuracy overall, AUC = .91, p <.001, 95% CI = [.85 .97]. Finally, we compared the AUCs of the various measures. The FCT and combined CIT did not differ, z = -0.19, p = .850. Combining the FCT and CIT featured a larger AUC than the FCT alone, z = -3.11, p = < .001, and combined CIT alone, z = -1.75, p = .040, further supporting the incremental validity of the FCT and CIT.

8. Discussion

We examined if an FCT that repeats information can be a diagnostic tool to detect concealed knowledge if only limited information about a crime is available. Results showed our novel FCT using only five details of the crime was as diagnostic as the well-established CIT.

A limitation of our design was that we did not counterbalance the sequence of CIT and FCT. However, we did so for a specific reason. We always started with the CIT, because we deemed it possible that presenting a subset of alternatives in the FCT would have made those answer alternatives meaningful for examinees without concealed knowledge, violating the assumption of equal probabilities in the CIT.

Unexpectedly, the distribution of the self-reported response strategies and their associated test scores differed from study 1. We observed that only one participant endorsed correct information similar to typical FCT studies such as Merckelbach et al. (2002) or Jelicic et al. (2004). However, given the misdirection manipulation as in study 1 we expected these participants to be more prevalent. A possible explanation for this finding could be that the FCT in study 2 also contained filler trials, unlike the one in study 1. Filler trials outnumbered trials with information from the mock crime 2 to 1 and this imbalance is obvious to participants with concealed knowledge. Therefore, participants may be more inclined to avoid correct information, because they only rarely encounter correct answers. Further investigation is needed.

9. General Discussion

We tested whether our novel FCT in which pieces of evidence are repeated, can be used in situations with insufficient evidence to construct a traditional FCT. In study 1, we provided a proof of concept by demonstrating that a repeated FCT can have diagnostic validity albeit in a highly controlled situation. In study 2, we validated the repeated FCT in a situation with insufficient evidence for a traditional FCT and compared it to the CIT, another well-established concealed knowledge detection tool based on psychophysiological indices. Again, our repeated FCT featured detection accuracies on par with previous studies, despite being based on less than half of the amount of information compared to, for example, Merckelbach et al. (2002), Giger et al. (2010), and Shaw et al. (2014).

If replicated, our findings that a diagnostic FCT can be produced with limited information has several potential implications. First, it means the FCT can be applied to more cases in practice, because fewer information is needed to create a valid test. Second, it makes it easier to integrate the FCT in a CIT examination. Typically, a CIT is conducted with around five pieces of evidence, so an investigator would have to collect much more evidence before an FCT can be added to the examination. Instead, a repeated FCT can be added without any additional requirements, and in line with previous research (Meijer et al., 2007), Study 2 demonstrates that combining the CIT and repeated FCT leads to a better detection accuracy. However, the selection of adequate pieces of information and safeguards against leakage become even more important for the FCT with repeated trials, because each leaked or unrelated piece of information affects a larger proportion of the test than in a traditional FCT. Therefore, measures to exclude leaked items such as previewing the questions and answers are integral to ensure the validity of the FCT examination.

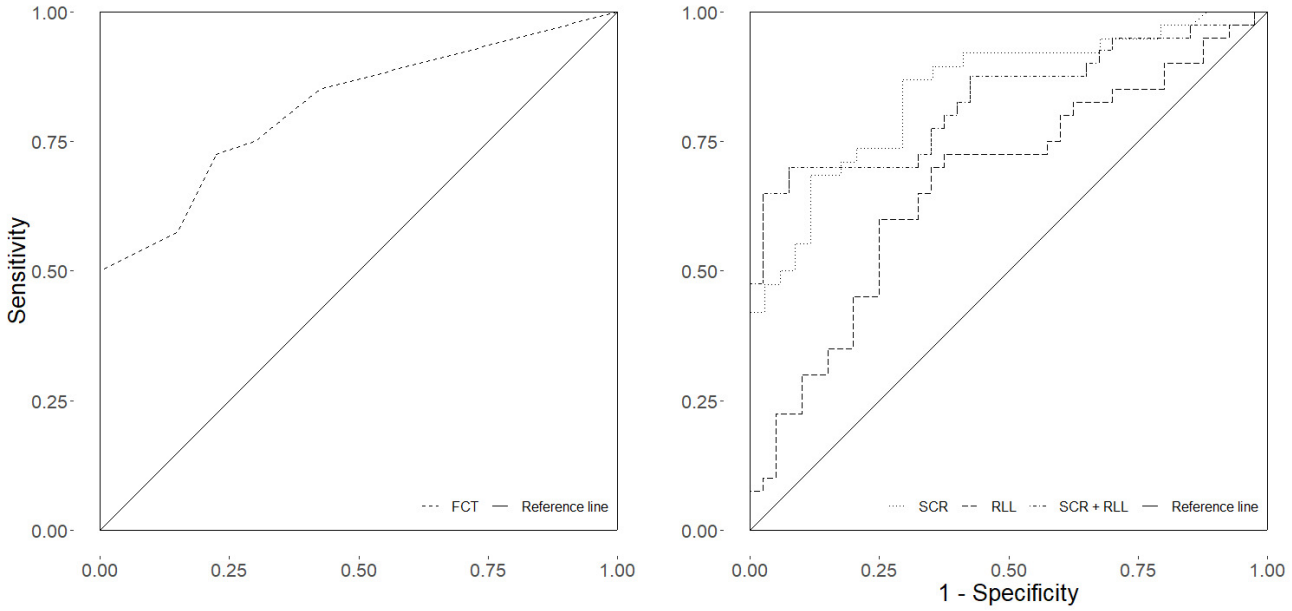


Figure 2. Receiver Operating Characteristic of the repeated FCT (left) and CIT measures: Skin Conductance Response, Respiration Line Length and their combination (right).

The findings from our two studies open up several new avenues for research. One important aspect is the distinction between test length and the information richness of the test. In a traditional FCT this distinction does not exist as each piece of information corresponds to one trial. In contrast, our novel FCT produces 20 trials from 5 pieces of information, allowing for a separation of test length and information richness. As such, it remains unclear how much the test length on the one hand, and information richness on the other, contribute to the detection accuracy of the FCT. Similarly, this distinction may also be crucial to understand the role of filler trials. Although filler trials are ignored by the examiner when computing how far the test score deviates from chance, they may still affect examinees' choosing behaviour during the test. For example, examinees may avoid more correct answers when filler trials outnumber the real FCT trials. Further research should also determine the contribution of departures from standard FCT procedure, such as the switch from a specific to generic question or the presence of a video camera as distractor, on the detection accuracy.

A second avenue worth looking into is the effect of increased test length on the diagnosticity of other criteria in the FCT. Orthey et al (2017) demonstrate that a considerable proportion of examinees with concealed knowledge understands how the FCT works, and rather than avoiding correct answers, randomize between correct and incorrect answers. This response strategy is not detected by the traditional underperformance criterion (e.g., Jelicic et al., 2004; Merckelbach et al., 2002) and other measures of randomness such as the number of alternations between correct and incorrect answers have been proposed as alternative criteria (Verschuere et al., 2008). However, as pointed out in Orthey et al. (2018) these measures require a large number of trials to elicit meaningful differences. Hence, they are unlikely to succeed in traditional FCTs, but a repeated FCT based on a medium amount of information could easily satisfy this requirement. A final aspect that deserves more attention is the role and effect of filler items. It is still unclear how the presence of filler items affects examinees' ability to see through the mechanism of the FCT. This is important, because the detection accuracy of the FCT is closely related to the examinees' understanding of the test's rationale (e.g., Orthey et al., 2017). Specifically, the frequency and the type of filler trials should be further investigated. In the original experiment Jelicic et al (2004) used questions about the scene of the crime, but not related to the act itself. In contrast, we constructed filler trials from the remaining incorrect CIT alternatives, which means that the filler trials, just like normal trials, contained, albeit incorrect, answer alternatives from crime related categories such as a weapon or method of entry.

Finally, we want to pre-emptively address a concern about the repeated presentation of information. The core problem of increasing test length by repeated presentation of information is that each trial must be independent from prior choices. Repeating a traditional FCT question violates this assumption, as without any new information the examinees best effort deduction should always result in the same choice, regardless of the actual underlying strategy; avoidance, intentional randomization, or genuine guessing. Here, we attempted to circumvent this problem by asking participants to make relative comparisons between two (unrelated) options. Selecting an option from a relative pair does not necessarily imply that this option is actually related to the crime (just more than the other alternative), meaning the next time the selected option is presented participants can change their response while being logically

consistent with their prior choices as a specific answer alternative can be perceived as more related to the crime than some, but not all, other answer alternatives. Hence, we argue that trial dependency did not occur in our novel procedure for three reasons. First, we specifically changed the design of the FCT so that despite the repeated presentation of individual pieces of information, each FCT trial was unique. That means the same combination of answer alternatives was never presented twice. The relative nature of the task, that is, indicating which answer is more related to the incident, combined with unique pairs, means prior choices should not be meaningful for subsequent trials. Second, if trial dependency had occurred, we would have expected to observe more extreme test scores for examinees without concealed knowledge. In contrast, our data in both experiments fall within the range of previous FCT studies that utilize a traditional test design. Third, we included a supplementary material that explores participants consistency throughout the test (Appendix A). In essence, we demonstrate that neither examinees with- or without concealed knowledge prioritize consistency with prior choices in their selection preference throughout the test.

In sum, over two experiments, we demonstrated that an FCT that repeats pieces of information has diagnostic accuracy similar to that of traditional FCTs, showing the FCT can be applied in situations with limited evidence.

Contributions

Contributed to conception and design: RO, EM

Contributed to acquisition of data: RO, EK

Contributed to data analysis and interpretation: RO, EM, NB

Contributed to writing and revising the manuscript: RO, EK, EM, NB

Competing Interests

There are no competing interests to declare.

Data Accessibility Statement

The datasets of study 1 and 2 are available for download at: https://osf.io/32hva/?view_only=cea2b312e1ee40018437e33a19f0b39d

Submitted: December 01, 2021 PDT, Accepted: July 25, 2022 PDT

References

Ben-Shakhar, G., & Elaad, E. (2002). Effects of questions' repetition and variation on the efficiency of the Guilty Knowledge Test: A reexamination. *Journal of Applied Psychology*, *87*(5), 972–977. https://doi.org/10.1037/0021-9010.87.5.972

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review. *Journal of Applied Psychology*, *88*(1), 131–151. https://doi.org/10.1037/0021-9010.88.1.131

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

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

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*(3), 837–845. https://doi.org/10.2307/2531595

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/10.1016/0887-6177(95)00042-9

Elaad, E., Ginton, A., & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology*, *71*(5), 757–767. https://doi.org/10.1037/0021-9010.77.5.757

Gati, I., & Ben-Shakhar, G. (1990). Novelty and significance in orientation and habituation: A feature-matching approach. *Journal of Experimental Psychology: General*, *119*(3), 251–263. https://doi.org/10.1037/0096-3445.119.3.251

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*(5), 440–463. https://doi.org/10.1080/15228932.2010.489875

Global Deception Research Team. (2006). A World of Lies. *Journal of Cross-Cultural Psychology*, *37*(1), 60–74. https://doi.org/10.1177/0022022105282295

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 31–49). Academic.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Jelicic, M., Merckelbach, H., & van Bergen, S. (2004). Symptom validity testing of feigned amnesia for a mock crime. *Archives of Clinical Neuropsychology*, *19*(4), 525–531. https://doi.org/10.1016/j.acn.2003.07.004

Klein Selle, N., Agari, N., & Ben-Shakhar, G. (2019). Hide or seek? Physiological responses reflect both the decision and the attempt to conceal information. *Psychological Science*, *30*(10), 1424–1433. https://doi.org/10.1177/0956797619864598

Lieblich, I., Naftali, G., Shmueli, J., & Kugelmass, S. (1974). Efficiency of GSR detection of information with repeated presentation of series of stimuli in two motivational states. *Journal of Applied Psychology*, *59*(1), 113–115. https://doi.org/10.1037/h0035781

Lukács, G. (2021). Prolonged response time concealed information test decreases probe-control differences but increases classification accuracy. *Journal of Applied Research in Memory and Cognition*, *11*(2), 188–199. https://doi.org/10.1016/j.jarmac.2021.08.008

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*(6), 385–388. https://doi.org/10.1037/h0046060

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, *44*(4), 258–262. https://doi.org/10.1037/h0044413

Lykken, D. T. (1974). Psychology and the lie detector industry. *American Psychologist*, *29*(10), 725–739. https://doi.org/10.1037/h0037441

Matsuda, I., & Ogawa, T. (2011). Improved method for calculating the respiratory line length in the concealed information test. *International Journal of Psychophysiology*, *81*(2), 65–71. https://doi.org/10.1016/j.ijpsycho.2011.06.002

Matsuda, I., Ogawa, T., & Tsuneoka, M. (2019). Broadening the use of the concealed information test in the field. *Frontiers in Psychiatry*, *10*, 24. https://doi.org/10.3389/fpsyt.2019.00024

Meijer, E. H., Klein Selle, N., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the concealed information test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology*, *51*(9), 879–904. https://doi.org/10.1111/psyp.12239

Meijer, E. H., Smulders, F. T., Johnston, J. E., & Merckelbach, H. (2007). Combining skin conductance and forced choice in the detection of concealed information. *Psychophysiology*, *44*(5), 814–822. https://doi.org/10.1111/j.1469-8986.2007.00543.x

Meijer, E. H., Verschuere, B., & Ben-Shakhar, G. (2011). 16 Practical guidelines for developing a CIT. In B. Verschuere, G. Ben-Shakhar, & E. H. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 293–302). Cambridge University Press. https://doi.org/10.1017/cbo9780511975196

Merckelbach, H., Hauer, B., & Rassin, E. (2002). Symptom validity testing of feigned dissociative amnesia: A simulation study. *Psychology, Crime & Law*, *8*(4), 311–318. https://doi.org/10.1080/10683160208401822

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

Orthey, R. (2019). *Response strategies of instructed malingerers during forced choice testing: New measures and criteria to detect concealed knowledge and feigned cognitive deficits*. https://doi.org/10.26481/dis.20190625ro

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

Orthey, R., Vrij, A., Meijer, E. H., Leal, S., & Blank, H. (2018). Resistance to coaching in forced-choice testing. *Applied Cognitive Psychology*, *32*(6), 693–700. https://doi.org/10.1002/acp.3443

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*. https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/july2003/podlesny.html

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Shaw, D. J., Vrij, A., Mann, S., Leal, S., & Hillman, J. (2014). The guilty adjustment: Response trends on the symptom validity test. *Legal and Criminological Psychology*, *19*(2), 240–254. https://doi.org/10.1111/j.2044-8333.2012.02070.x

Siegel, S. (1956). *Nonparametric statistics for the behavioural sciences*. 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 Oorsouw, K., & Merckelbach, H. (2010). Detecting malingered memory problems in the civil and criminal arena. *Legal and Criminological Psychology*, *15*(1), 97–114. https://doi.org/10.1348/135532509x451304

Verschuere, B., & Crombez, G. (2008). Déjà vu! The effect of previewing test items on the validity of the Concealed Information polygraph Test. *Psychology, Crime & Law*, *14*(4), 287–297. https://doi.org/10.1080/10683160701786407

Verschuere, B., Meijer, E., & Crombez, G. (2008). Symptom Validity Testing for the detection of simulated amnesia: Not robust to coaching. *Psychology, Crime & Law*, *14*(6), 523–528. 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*(1), 65–72. https://doi.org/10.1037/h0032060

Appendix A

In a traditional FCT a factual question about the crime is presented with two related answer alternatives, for example "gun" and "knife" denoted here in abstract terms: 1 & A. In our repeated FCT we transform the specific questions to a relative one, "Which of the answers is more related to the crime?" and pair each correct answer with all incorrect answers of the other questions, for example 1 & C and 1 & B... Consequently, the total number of possible trials is increased significantly.

Here we explore how consistent participants are throughout the test. If participants do treat trial with repeated information as dependent, we would expect them to be consistent in their choice over all presentations of that answer throughout the test.

Consistency

In experiment 2, the repeated FCT featured 20 critical and 42 filler trials that were created from 5 and 7 respective traditional FCT questions. That means according to our method each piece of evidence from critical trials is presented four- and each piece of evidence from filler items is presented six times throughout the entire test. We computed a consistency rating for each piece of information, indicating the most prevalent choice over all presentations.

A critical item is presented 4 times, so the consistency can assume three different values:

- 50% = the item was selected and avoided two times
- 75% = the item was selected once and avoided thrice or selected thrice and avoided once
- 100% = the item was either selected or avoided four times

A filler item is presented 6 times, so the consistency can assume four different values:

- 50% = the item was selected and avoided three times
- 66.6% = the item was selected four times and avoided two times or selected two times and avoided four times
- 83.3% = the item was selected five times and avoided once or avoided five times and selected once
- 100% = the item was selected or avoided six times

First, we will examine the distribution of consistencies over critical and filler items per piece of information (see next page) for participants without concealed information.

- We can observe that the distribution of scores is not dominated by 100% consistency.
- We can observe the same pattern for critical and filler items. (By definition, innocent participants should not be able to distinguish critical and filler items and hence respond in the same manner).

Next, we will look at the distribution of consistency over critical and filler items for participants with concealed knowledge.

It is important to realize that not all malingerers follow the same response strategy. In a typical sample around half of malingerers simply avoid correct answers and produce underperformance. Most of the remaining half typically follows a response strategy to mix correct and incorrect answers, leading to test scores that fall within chance levels and finally, a small proportion of this group does follow the test instructions and endorses correct answers. Hence, we must consider these response strategies when evaluating the consistency.

Concretely, participants who avoid or endorse correct answers as part of their response strategy would produce high consistency ratings, irrespective of whether they perceived the trials as dependent or independent.

Below we display the distribution of consistency scores for critical and filler items for participants with concealed knowledge. We note the following:

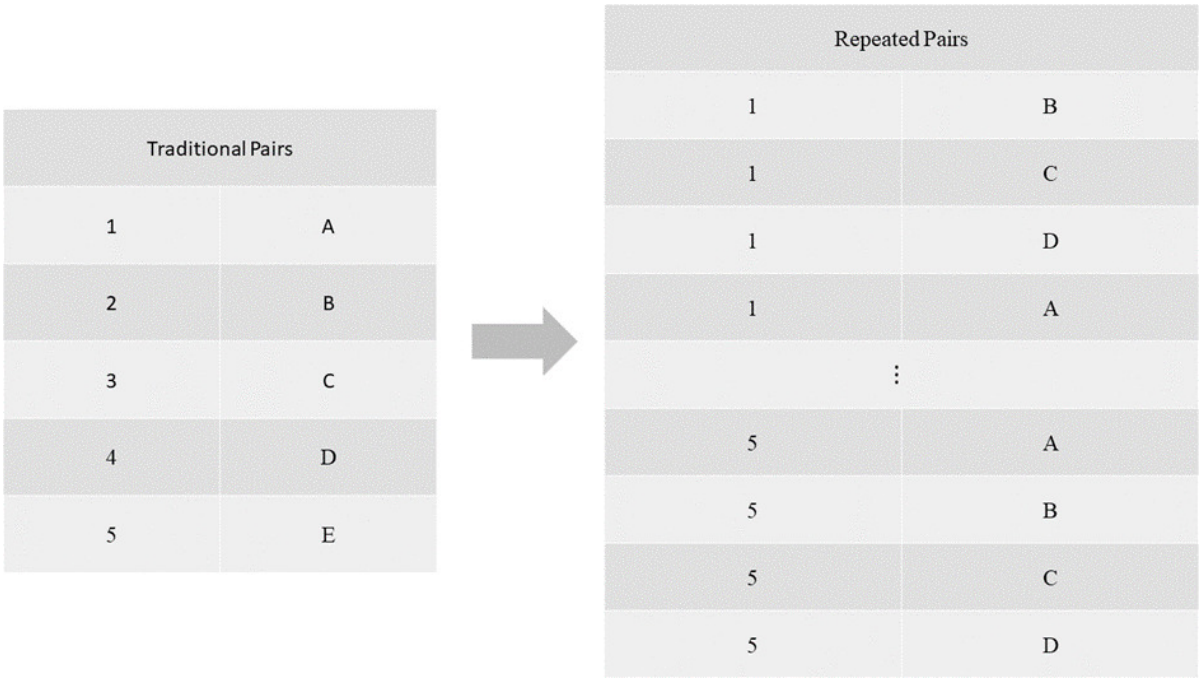
- There is a clear dominance of 100% consistency scores for critical items. However, this dominance can be attributed to the over- and underperformance response strategy.
- Malingerers who randomize between correct and incorrect answers do not do so consistently, in other words always endorsing/avoiding the same answer.
- There is no dominance of 100% consistency scores for filler items. Not even for malingerers following the over- or underperformance strategy. Further supporting that the high consistency found for critical items in these groups is a consequence of their response strategy rather than trial dependency.

Finally, we must examine the overall response pattern (see next page). While there is no dominant pattern for complete consistency on the group level per piece of information, it is still possible that a small subsample of malingerers and innocents performed perfectly consistent over the entirety of the test. To examine the consistency over all items, we averaged the scores for critical and filler within individuals and plotted the distribution of their scores.

We note the following:

- The scores for participants without concealed information are distributed over the entire range of possible values, except for the upper end of the scale. There were no participants who featured high consistencies over all pieces of information combined.
- For participants with concealed information who randomised between correct and incorrect answers we observe the same pattern as for innocent participants.
- Only participants with concealed information who follow over- or underperformance strategies predominantly produce a high consistency over all items. Importantly, this pattern only applies to critical items and not to filler items. This again supports the notion that high consistency ratings were a consequence of their response strategy rather than perceived trial dependency.

In sum, we demonstrate that participants without concealed knowledge are not consistent over individual pieces of information or over the entirety of the test. Participants with concealed knowledge who randomised between correct and incorrect answers follow the same trend. High consistency is only displayed by participants with concealed



Traditional Pairs	
1	A
2	B
3	C
4	D
5	E

Repeated Pairs	
1	B
1	C
1	D
1	A
⋮	
5	A
5	B
5	C
5	D

Figure A1.

knowledge who either avoid or endorse correct information on purpose and notably only for critical but not filler trials. This suggests that their high consistency is a product of their response strategy rather than trial dependency. However, their consistent choosing pattern is better explained by their chosen response strategy rather than perceived trial dependency, because this behaviour only occurs for critical and not filler items.

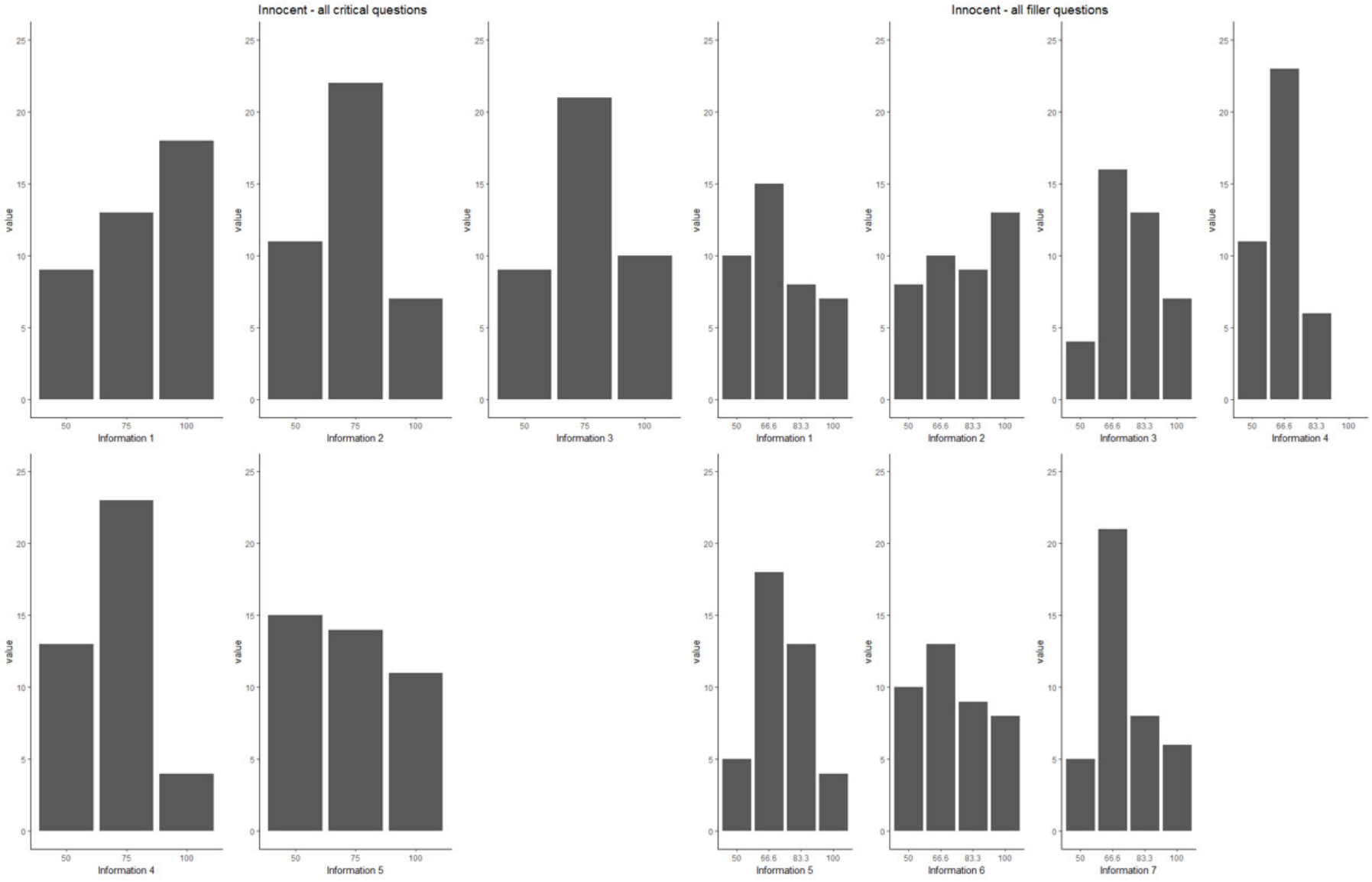


Figure A2.

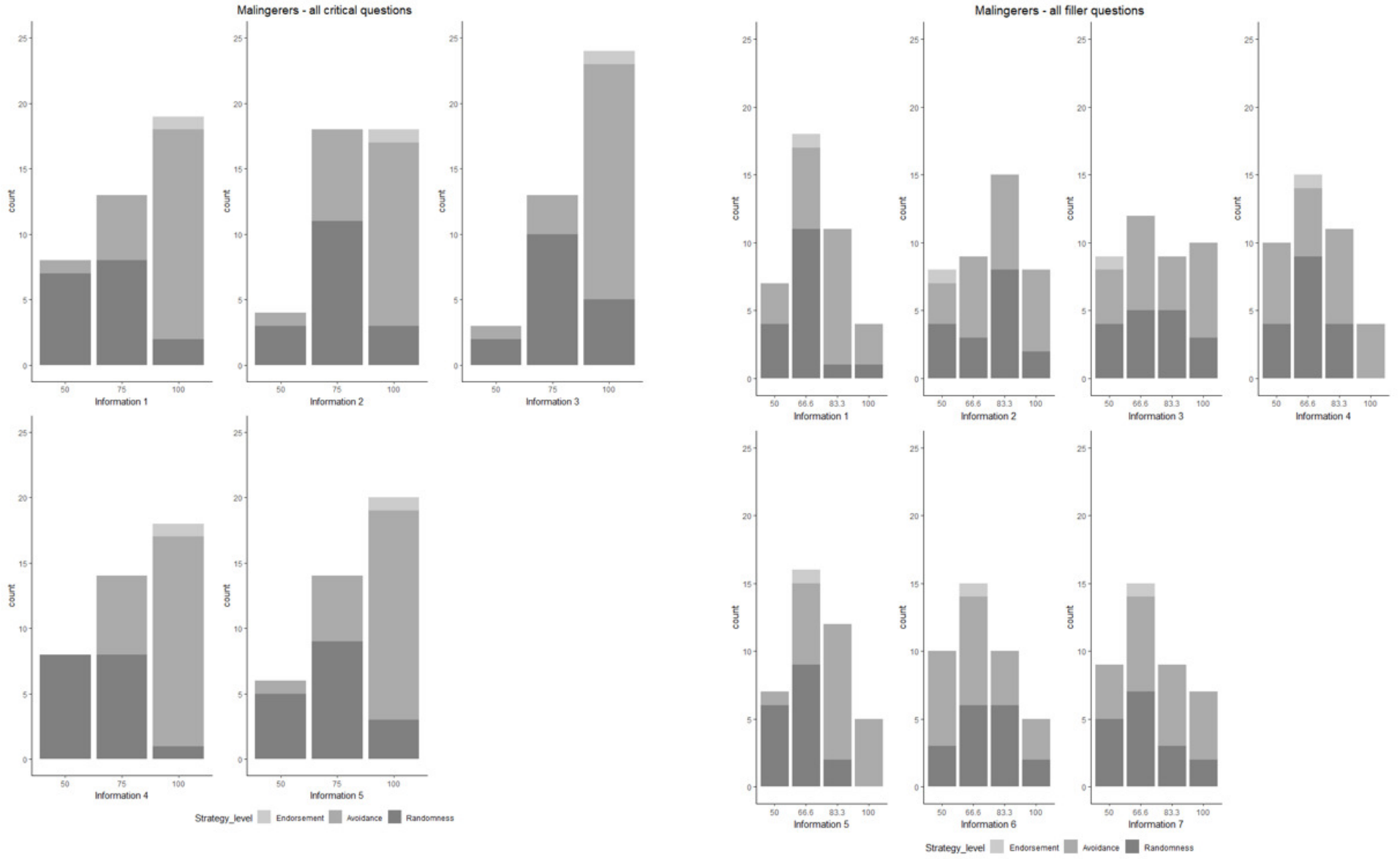


Figure A3.

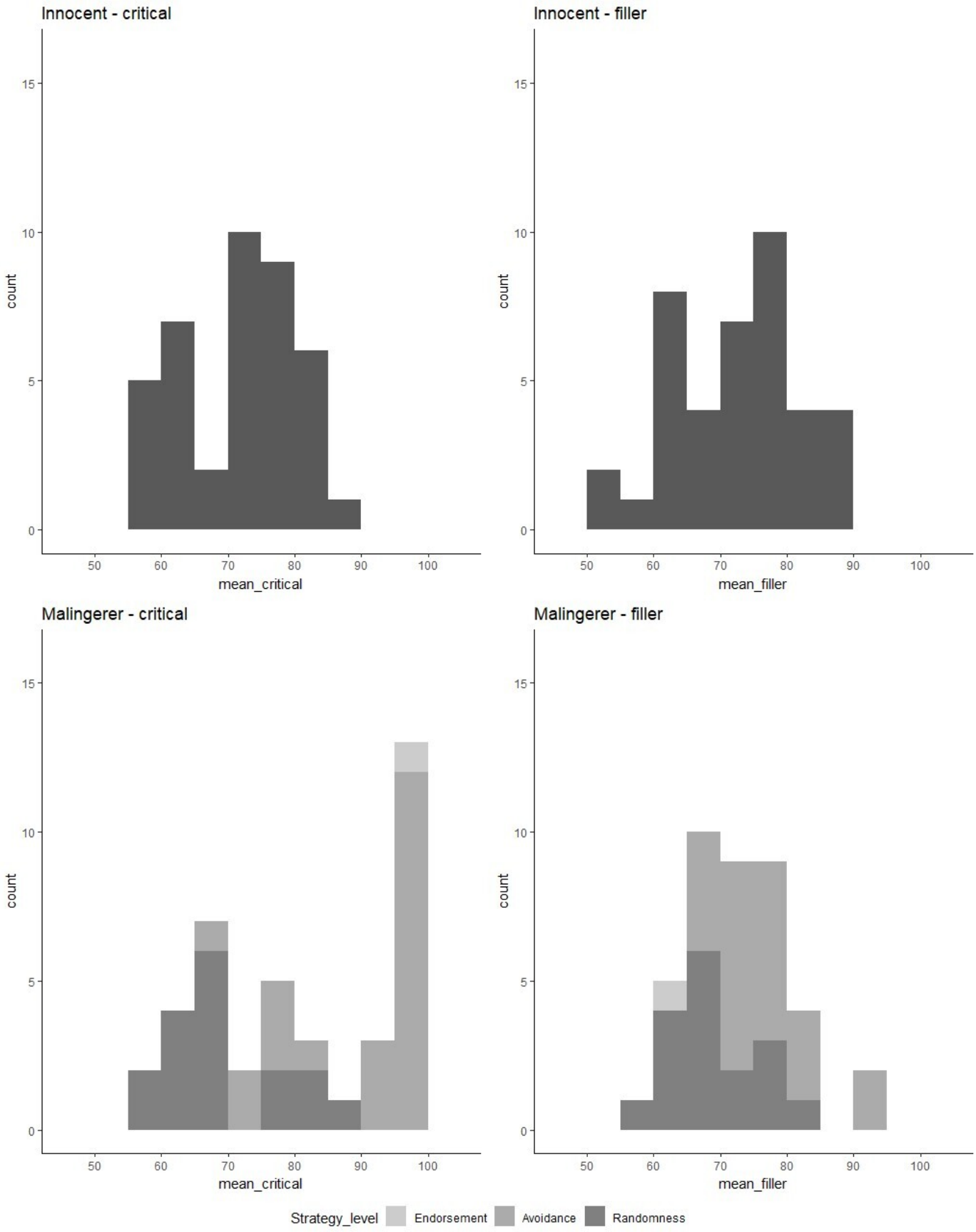


Figure A4. Average consistency over conditions

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/37483-how-to-detect-concealed-crime-knowledge-in-situations-with-little-information-using-the-forced-choice-test/attachment/95895.docx?auth_token=vJKf9qdCPQkl4nt68iin