**Peer Review and Communication History**

**MS Title**: How to Detect Concealed Crime Knowledge in Situations with Little Information Using the Forced Choice Test

**Author Names**: Robin Orthey, Ewout Meijer, Emmeke Kooistra, Nick J Broers

**Submitted:** Dec 2, 2021

**Editor First Decision**: Revise & Resubmit

Jan 19, 2022

Dear Robin Orthey,

I have now received all reviews of your manuscript, “How to Detect Concealed Crime Knowledge in Situations with little Information using the Forced Choice Test” from qualified researchers. All three reviewers agreed on the high quality of your manuscript and its substantial contribution. However, reviewers also provided some suggestions for a revision and raised some points which should be included in the discussion of your manuscript. I therefore strongly encourage you to submit a revised version for further consideration at Collabra: Psychology.

I think this is a promising manuscript and I very much look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at [editorialoffice@collabra.org](mailto:editorialoffice@collabra.org).

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,  
Martin Schultze

**Reviewer 1**

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The present study aims to extend the real-life applicability of the Forced Choice Test (FCT) deception detection method by allowing its use with lesser number of available incriminating items. This makes sense, and the study seems to me methodologically well conducted; I have no objection against publication. I only have some minor comments, listed below.

* I realize that this comment is not specifically about the present manuscript, but, generally, the motivation behind FCT research is not entirely clear to me, given that there is already a well-established and intensely research alternative in the CIT. Of course, I do not mean to discourage developing or advancing alternatives, but I think there should be some fair reasoning behind investing into such efforts, at least “in-principle”, that is, if not proven, at least assumed with reasonable plausibility. As it stands, the FCT seems to have only disadvantages as compared to the CIT. Not only does it need more items (even with the present improvements), but I strongly suspect that its simplicity makes it very susceptible to faking. All in all, I would have been more enthusiastic about a CIT study with a similar aim (getting more out of fewer items).

p. 3: As I understand, the detection rates apply to “guilty” groups only. Without an innocent control group, or at least without really well justified and verified cutoff (5% FPR seems arbitrary without further information), this seems quite meaningless. The reported detection rates around 50% imply chance level accuracy (i.e., that the method is completely useless), if I were to assume that the detection rate is similar (i.e., also around 50%) in an innocent sample.

p. 6: The number of participants should be reported per group, not just altogether. The sample size in that case seems rather low.

p. 7: “participants who understand the FCT`s rationale are more successful at avoiding detection” I doubt that, if the FCT were ever applied in real life, examiners would be (legally) allowed, let alone able, to conceal the rationale of the test from the examinees.

p. 8: How was the VR scenario constructed (what software, etc.)? Some supplementary images (screenshots) would be helpful. Also: is this material (the VR program) shared openly/publicly?

p. 11: What would make sense is to statistically compare the AUCs, but I do realize it is unlikely to yield significance differences with these rather small samples per group.

p. 14: The autonomic response based CIT is of course just one kind of CIT, and even that has many variations (depending on what responses are measured, how it’s evaluated, etc.). I think this should be briefly clarified or noted.

p. 14: “Evidence further suggests that combining he FCT and CIT has a better diagnostic accuracy than either test alone.” What evidence? I’d be curious, because this may be one good reason for FCT research, but there is no direct citation.

p. 15: Again I wonder whether this material is available publicly. I realize that the Authors intended to mask their identity, but an anonymized OSF repository would have been ideal to share for review, or otherwise the material could have been simply attached anonymously via the journal’s submission system.

p. 21: Again, the statistical comparison of AUCs seems desirable, but again I know that with these sample sizes it will yield neither significant differences, nor narrow confidence intervals to estimate noninferiority.

* I’d advise proofreading, since there are a few small issues or at least strange choices in notation and grammar throughout the paper. (E.g., already in the abstract, the sentence with “[…] trial is unique the properties […]”, perhaps not strictly incorrect, but would be much easier to read including a comma like “[…] trial is unique, the properties […]”. Another example is the repeated mixed use of " ` " instead of " ’ ". Or that abbreviations such as “e.g.” are not supposed to be used outside parenthesis, at least according to APA style.)

**Reviewer 2**

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The main challenge of the Forced Choice Test (FCT) is that it often requires more evidence than what is typically available from a crime. The current study examines whether this challenge can be circumvented by repeating individual pieces of information until the necessary test length is achieved. Study 1 demonstrates that a repeated FCT has diagnostic validity. Study 2 extends this finding to a situation with insufficient evidence for a traditional FCT.

Overall, I think the idea for the study is interesting, the study design is straightforward, the sample size is sufficiently large, and the analysis is done according to widely accepted standards. In sum, I have no major concerns about the manuscript, but I do think that there are several points that need to be addressed:

**Introduction**  
• On p. 3, you write: “The only exception is study 2 from Meijer et al (2007), but unlike all other studies a community sample as opposed to a student sample, was used.” I found this a very strange statement to read. How could the use of a community sample explain that in this particular study more trials did not result in higher detection accuracy? Please explain or rephrase.

**Method**  
• P. 6 & 14: As you did not do any power analyses, I wonder how you determined the size of your participant sample.  
• On p. 14, you write: “we subjected half of our participants to a mock crime” Sentences like this make it sound as if participants in your study had to execute a mock crime. However, I wouldn’t call reading about an imaginary terror organization a mock crime. Please rephrase or explain why you would think your method should be called a mock crime.

**Results**  
• P. 10 & 20: You have a lot of level 0 strategies in study 1, but not in study 2. You suggest that this could be due to the fact that you used filler items in study 2 and not in study 1. An explanation that would make even more sense to me would be that it may not have been entirely clear to some participants that they needed to hide their crime knowledge, or that they were not motivated to do so because you gave no reward in this study and you did in study 2. In the original study about strategy and misdirection by Orthey, Vrij, Leal, and Blank (2017), the authors also found almost no participants who used strategy 0 to conceal their knowledge. However, you found 27.50% and 28.21%, which might be unproportionally high (and might also explain your extremely high accuracies, see below).  
• P. 11: You find very high accuracy scores for the FCT. As you state yourself in your paper, these high accuracies could be due to counting both under- and overperformance as diagnostic of having concealed information. However, as this way of calculating is less conservative than the classic way of calculating, I would like to see what you would find if you would calculate it in the classic way as well. Especially because statements like “The accuracy of our novel procedure is in line with previous experiments” make no sense if you calculated your variable in a different way than the previous experiments you want to compare with.

**Discussion**  
• On p. 25, you state that measures of randomness such as the number of alternations between correct and incorrect answers have been proposed as an alternative way to detect concealed knowledge with the FCT. Traditional FCT’s are usually too short to apply this test, but you state that a repeated FCT could be a solution to this problem as it “could easily satisfy this requirement.” As you have all the variables to test this hypothesis, I think it would make your discussion much stronger if you could back up these statements with actual numbers that prove this idea.

**General remarks**  
• One major concern I have with the repeated FCT is that the leakage problem might become more pronounced in it. If an innocent suspect has some knowledge about a crime (e.g., what the murder weapon was), then instead of looking guilty on only one trial as in the traditional FCT or CIT, the suspect might now look guilty on multiple trials because all the information in the repeated FCT is intermingled with each other. For example, in a crime where only 5 pieces of information are available to the police, 8 out of 20 trials contain stimuli related to the murder weapon. In this way, a knowledgeable innocent suspect would appear guilty on almost half of the trials. This problem becomes even larger if fewer items are being repeated. When there are only 4 pieces of information available, 6 out of 12 trials contain information about 1 item, when there are only 3 pieces of information, 4 out of 6 trials contain information about the same item, etc. I’m curious to hear what you think about this concern.

**Reviewer 3**

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

This is an interesting and very well-written manuscript examining the validity of a new method of developing a Forced Choice Test (FCT) in order to detect concealed crime knowledge. I enjoyed reading it. I think both studies are based on carefully planned designs and analyses. The text is both, easy to read and easy to follow.

My only suggestions relate to the role of the filler trials that were added in study 2. Adding these item sets actually changes the rationale of the FCT, since the test strategy is based on the idea that respondents try to conceal their knowledge. Therefore, the new FCT in study 1 does not represent the same test as the new FCT in study 2. Perhaps this new FCT test strategy in study 2 works better than the FCT test strategy in study 1 due to the filler items and their effect on answer strategies (higher frequency of strategy 1; comparing Table 1 with Table 3). Therefore, it would be helpful if you elaborated more on the role of the filler items as well as the practical implications of this finding e.g., how to construct adequate answer alternatives that would lead to the same test performance as the FCT you present in your study. As it can be read in the appendix page 33, line 28, filler trials are presented a higher number of times than the critical trials, therefore the quality of those items affect the test performance substantially.

Minor points

Page 21, line 13: the sentence is missing a full stop

Page 26, line 7: I think it should be “fall” instead of “falls”, since data is plural.

Page 5, line 10 and 11: it would be helpful for the reader, if you mentioned that in the first FCT the proportion of each crime-related answer alternative (e.g., “knife”) remains the same, even though it is paired with multiple non-related answers.

**Author Response**  
Feb 24, 2022

*Dear Editor and reviewers,*

*Thank you very much for your time and effort in reviewing our manuscript. Please find our answers to your comments in italics below the relevant comments. Additionally, when we quote sections from the manuscript, changes are highlighted in bold.*

Reviewer 1

The present study aims to extend the real-life applicability of the Forced Choice Test (FCT) deception detection method by allowing its use with lesser number of available incriminating items. This makes sense, and the study seems to me methodologically well conducted; I have no objection against publication. I only have some minor comments, listed below.

I realize that this comment is not specifically about the present manuscript, but, generally, the motivation behind FCT research is not entirely clear to me, given that there is already a well-established and intensely research alternative in the CIT. Of course, I do not mean to discourage developing or advancing alternatives, but I think there should be some fair reasoning behind investing into such efforts, at least “in-principle”, that is, if not proven, at least assumed with reasonable plausibility. As it stands, the FCT seems to have only disadvantages as compared to the CIT. Not only does it need more items (even with the present improvements), but I strongly suspect that its simplicity makes it very susceptible to faking. All in all, I would have been more enthusiastic about a CIT study with a similar aim (getting more out of fewer items).

*We do not believe things are as simple as that the FCT is obsolete now that the CIT is around. The FCT has a nearly 50-year history and is a generally accepted tool in Neuropsychology. Not having to rely on technology or extensive training, it has several advantages over the CIT: it is cheap, fast, and easy to use. Rather than one replacing the other, we believe they are complementary as they rely on different psychological mechanisms, which is supported by the incremental validity shown in experiment 2.*

p. 3: As I understand, the detection rates apply to “guilty” groups only. Without an innocent control group, or at least without really well justified and verified cutoff (5% FPR seems arbitrary without further information), this seems quite meaningless. The reported detection rates around 50% imply chance level accuracy (i.e., that the method is completely useless), if I were to assume that the detection rate is similar (i.e., also around 50%) in an innocent sample.

*There seems to be a misunderstanding here. On page 3 we summarize the detection rates as reported in the literature. It is true that many, especially older, studies only report the detection rates for guilty participants. These detection rates are still informative, because performance of innocent participants can be assumed to follow the binomial distribution. Hence, the cut-off points used by the authors can be expected to result in a 5% false positive rate, corresponding to a 95% correct classification of innocent participants. As a result, a, for example, reported detection of 50% of the guilty participants is not chance level. Rather, it corresponds to a 72.5% overall accuracy ((50+95)/2). To clarify this we rewrote the beginning of the second paragraph on p.3.*

*Of course, we fully agree with the comment that control groups are desirable and that is why we have included them in all our studies. Still, as innocent’s behaviour conforms to the chance distribution in a correctly constructed FCT we do believe these studies are worth mentioning.*

p. 6: The number of participants should be reported per group, not just altogether. The sample size in that case seems rather low.

*We have specified the exact distribution of participants for study 1 in the method sections:*

*“Examinees were unaware of the total number of questions and were assigned to either a traditional FCT or a 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.****”*

*We have done the same for study 2:*

*“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.”*

p. 7: “participants who understand the FCT`s rationale are more successful at avoiding detection” I doubt that, if the FCT were ever applied in real life, examiners would be (legally) allowed, let alone able, to conceal the rationale of the test from the examinees.

*The FCT is commonly used in neuropsychological test batteries, including to test for amnesia for a crime (Denney, 1996; Jelicic, 2018), and the results often play a role in court proceedings. The exact rationale of the FCT is not explained in the informed consent, and apparently this meets both the professional code of conduct as well as legal requirements. If the reviewer can supply us with a reference or ruling that describes his/her point of view we will happily discuss that in our revision.*

*Secondly, practical and experimental studies demonstrate that even if the rationale is explained the FCT can still perform above chance level (e.g., Denney, 1996; Jelicic et al., 2004; Giger et al., 2010; Merckelbach et al. 2002, Meijer et al., 2007; Orthey et al., 2017, 2019, etc…). However, this discussion is beyond the scope of our manuscript.*

p. 8: How was the VR scenario constructed (what software, etc.)? Some supplementary images (screenshots) would be helpful. Also: is this material (the VR program) shared openly/publicly?

*The scenario was constructed in the unity video game engine. We have updated the text as follows:*

*“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/).”***

*As part of the resubmission we provide some screenshots of the scenario that we will make available on the OSF page alongside the data as described in the data availability statement.*

*Unfortunately, the VR scenario itself is not publicly available. It was retired and we no longer hold a licence for the software. We would like to add that we don’t think the exact mock crime itself is essential to the experimental design and that it can be replaced by any mock crime / concealed knowledge paradigm in use today.*

p. 11: What would make sense is to statistically compare the AUCs, but I do realize it is unlikely to yield significance differences with these rather small samples per group.

*We have included comparisons for unpaired AUCs between the traditional FCT and repeated FCT with 100 trials as well as the traditional FCT and repeated FCT with 20 trials.*

*We made the following addition to the method section:*

*“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, Turck, Hainard, Liscaek, Sanchez, & Muller, 2011)”***

*And result section:*

*“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.*** *“*

p. 14: The autonomic response based CIT is of course just one kind of CIT, and even that has many variations (depending on what responses are measured, how it’s evaluated, etc.). I think this should be briefly clarified or noted.

*We agree, there are many facets to the CIT. We have chosen for the standard application as most commonly used in research and practice in Japan, as the focus of our studies was the FCT and not the CIT. We have clarified this as follows:*

*“The CIT examination followed the standard protocol in the literature (e.g., Matsuda, Ogawa, & Tsuneoka, 2019)* ***as it resembles best the current application of the CIT by the police in Japan.*** *We formulated five questions about critical****”***

p. 14: “Evidence further suggests that combining he FCT and CIT has a better diagnostic accuracy than either test alone.” What evidence? I’d be curious, because this may be one good reason for FCT research, but there is no direct citation.

*We are very thankful for this comment. This was a simple oversight. We have modified our manuscript as follows:*

*“Evidence further suggests that combining he 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 examinees without concealed knowledge and a 57% detection rate for examinees with concealed knowledge. In study 2, the combination of CIT and FCT scores produced a 88% detection rate for examinees with concealed knowledge. In both studies the combined score exceeded the individual CIT and FCT scores.”***

p. 15: Again I wonder whether this material is available publicly. I realize that the Authors intended to mask their identity, but an anonymized OSF repository would have been ideal to share for review, or otherwise the material could have been simply attached anonymously via the journal’s submission system.

*The materials contain elements that can be used under the fair use principle for non-commercial research purposes, but we do not have the right to share them. We have described the materials in detail as well as the procedure how we validated them as is the norm in this field.*

p. 21: Again, the statistical comparison of AUCs seems desirable, but again I know that with these sample sizes it will yield neither significant differences, nor narrow confidence intervals to estimate noninferiority.

*We made the following change to the method section:*

*“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 ≥ 2.58; z ≥ 1.96).”*

*We updated the results section as follows:*

*“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.*** *”*

*We also updated Table 4 to now include the AUC of the FCT & CIT combined measure.*

* I’d advise proofreading, since there are a few small issues or at least strange choices in notation and grammar throughout the paper. (E.g., already in the abstract, the sentence with “[…] trial is unique the properties […]”, perhaps not strictly incorrect, but would be much easier to read including a comma like “[…] trial is unique, the properties […]”. Another example is the repeated mixed use of " ` " instead of " ’ ". Or that abbreviations such as “e.g.” are not supposed to be used outside parenthesis, at least according to APA style.)

*We made these changes and improved readability throughout the entire manuscript without changing the content itself. All changes are marked in the track changes.*

Reviewer 2

The main challenge of the Forced Choice Test (FCT) is that it often requires more evidence than what is typically available from a crime. The current study examines whether this challenge can be circumvented by repeating individual pieces of information until the necessary test length is achieved. Study 1 demonstrates that a repeated FCT has diagnostic validity. Study 2 extends this finding to a situation with insufficient evidence for a traditional FCT.

Overall, I think the idea for the study is interesting, the study design is straightforward, the sample size is sufficiently large, and the analysis is done according to widely accepted standards. In sum, I have no major concerns about the manuscript, but I do think that there are several points that need to be addressed:

**Introduction**  
• On p. 3, you write: “The only exception is study 2 from Meijer et al (2007), but unlike all other studies a community sample as opposed to a student sample, was used.” I found this a very strange statement to read. How could the use of a community sample explain that in this particular study more trials did not result in higher detection accuracy? Please explain or rephrase.

*There is a slight misunderstanding here. We are not suggesting that using a community sample removed the (suggested) effect that longer test are more accurate. Rather, we note that the use of another population may also explain why the detection accuracy is high, despite the test being short. We have moved the passage in question to the following paragraph to improve readability modified the manuscript as follows:*

*"* *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, Naftali, Shmueli, and Kugelmass, 1974; Lukacs, 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 minimal test length, but 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. “***

**Method**  
• P. 6 & 14: As you did not do any power analyses, I wonder how you determined the size of your participant sample.

*Our main measures of interest are sensitivity, specificity, and the Area Under the Curve (AUC). All three measures are effect sizes, which means that the larger the sample grows the more precise our estimates become. Therefore, we followed the strategy maximizing precision given resource constraints (see Lakens (Preprint) Sample size justification). The only limitations we placed on ourselves were that we aimed for equal group sizes and that our sample should be at least as large as most studies published in this field.*

• On p. 14, you write: “we subjected half of our participants to a mock crime” Sentences like this make it sound as if participants in your study had to execute a mock crime. However, I wouldn’t call reading about an imaginary terror organization a mock crime. Please rephrase or explain why you would think your method should be called a mock crime.

*We have changed the section to avoid confusion. It now reads:*

*“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.****”*

**Results**  
• P. 10 & 20: You have a lot of level 0 strategies in study 1, but not in study 2. You suggest that this could be due to the fact that you used filler items in study 2 and not in study 1. An explanation that would make even more sense to me would be that it may not have been entirely clear to some participants that they needed to hide their crime knowledge, or that they were not motivated to do so because you gave no reward in this study and you did in study 2. In the original study about strategy and misdirection by Orthey, Vrij, Leal, and Blank (2017), the authors also found almost no participants who used strategy 0 to conceal their knowledge. However, you found 27.50% and 28.21%, which might be unproportionally high (and might also explain your extremely high accuracies, see below).

*Thank you very much for this comment. We understand why this misunderstanding occurred. The way strategies were measured was updated in Orthey et al. (2018), to align the strategy scores better to FCT scores. This update only affects one particular group:*

*In Orthey et al (2017) some participants were misdirected with a fake polygraph machine and consequently performed countermeasures against the polygraph while endorsing correct answers. According to the original definition this behaviour was classified as a level 1 strategy, a response towards the test instructions. However, since Orthey et al. 2018 classifications are made with closer reference to the selection behaviour rather than the intention behind it. So, if the definition of Orthey et al 2018 had been used in Orthey et al 2017 participants who reported response strategies such as controlling their demeanour or polygraph countermeasures would likely be classified as level 0 strategies as opposed to level 1. The description of strategy levels in the manuscript already reflected the most up to date definition (Orhtey et al 2018):*

*Page 8 lines 18-20: “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.”*

*, but we cited the wrong reference (Orthey et al 2017). Still, the implication of Orthey et al (2017) that the misdirection manipulation can promote endorsement of correct answers remains unchanged and was observed in our study as well. That means our distribution of FCT test scores is in line with those from Orthey et al (2017).* ***Hence, we updated the reference in the description of strategy levels to Orthey et al. 2018***

*To address the comments about the alternative explanation that money may be a deciding factor or that participants simply did not understand the experiment; First, we debriefed participants after each study and if we noticed that the test instructions were misunderstood those participants are excluded. We excluded none for this reason. Secondly, it is true that there was no financial incentive in study 1 and there was one in study 2, but in our experience this does not affect the strategy selection. All response strategies tend to take the same amount of time, so participants do not finish faster if they do a poor job to get the experiment over with. Also, for example Orthey et al 2017 for example only offered a raffle for a 50 pound voucher and did not pay everyone.*

*Please see our response to reviewer three about the strategy levels in study 2.*

• P. 11: You find very high accuracy scores for the FCT. As you state yourself in your paper, these high accuracies could be due to counting both under- and overperformance as diagnostic of having concealed information. However, as this way of calculating is less conservative than the classic way of calculating, I would like to see what you would find if you would calculate it in the classic way as well. Especially because statements like “The accuracy of our novel procedure is in line with previous experiments” make no sense if you calculated your variable in a different way than the previous experiments you want to compare with.

*Thank you for this comment. Study 1 had a misdirection manipulation and we used a bidirectional criterion. Only Orthey et al 2017 did the same and our reported detection accuracy is very similar with sensitivity above 50% at very high specific. So, strictly speaking our findings to match the only other study that is reasonably comparable.*

*We have relativised our discussion as follows:*

*“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 of Meijer et al. (2007), or Shaw et al. (2014), Jelicic et al. (2004) and Meijer et al. (Study 2, 2007)* ***both types of FCT were able to detect meaningful proportions of examinees 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. ”*

*We also recalculated the FCT diagnostic data for the unidirectional criterion in study 2. For the unidirectional criterion we used 1.65 and 2.33 as respective 5% and 1% cut offs. We found a sensitivity of 47.5% and 100% specificity for the 1% cut-off. For the 5% cut-off we found a 55% sensitivity and a 92.5% specificity. The AUC was .81 and statistically significant from zero (95% CI = [.70 - .91]). The results here are essentially the same as we already report, and we are not convinced that adding all these analyses improved readability, and think it may distract from the core conclusions that are based on the findings in study 2. We would like to note, however, that both data sets will be available after publication, so anyone can recalculate accuracy according to their specific needs.*

**Discussion**  
• On p. 25, you state that measures of randomness such as the number of alternations between correct and incorrect answers have been proposed as an alternative way to detect concealed knowledge with the FCT. Traditional FCT’s are usually too short to apply this test, but you state that a repeated FCT could be a solution to this problem as it “could easily satisfy this requirement.” As you have all the variables to test this hypothesis, I think it would make your discussion much stronger if you could back up these statements with actual numbers that prove this idea.

*We have run the analyses. The AUC for the traditional FCT did not differ from chance, AUC = .56, p = .307, 95%CI = [.44 - .69]. The AUC for the repeated FCT with 100 trials also did not significantly differ from chance, AUC = .49, p = .854, 95%CI = [.38 - .64].*

*We are, however, hesitant to include these findings for two reasons. First, the repeated FCT in study 2 is a full application of our novel method, but with only 20 trials it is too short to test this question.*

*That leaves only the repeated FCT from study 1. There are two considerations why the repeated FCT from study is suboptimal to answer this question. First, as pointed out in our manuscript the repeated FCT from study 1 best serves as a proof of concept if information can at all be repeated. To put it into context with a traditional FCT we had to make compromises. The repeated FCT here also contains 20 unique pieces of information, each repeated 5 times. Hence, although we have achieved a test length of interest, we used more pieces of information than we should have had, had we followed the procedure to the latter as in study 2. So, the repeated FCT in that condition neither fully represents the traditional or repeated FCT. Second, the runs test only detects intentional randomisation strategies, meaning our sample for examinees with concealed knowledge is effectively halved. This is not an issue for the underperformance criterion, because the effect is very large. However, the runs test relies on the human inability to reproduce randomness, which has a much smaller effect size (see Wagenaar 1972). Therefore, we believe that any runs test conducted here would not be meaningful and needed to be investigated on its own further.*

**General remarks**

One major concern I have with the repeated FCT is that the leakage problem might become more pronounced in it. If an innocent suspect has some knowledge about a crime (e.g., what the murder weapon was), then instead of looking guilty on only one trial as in the traditional FCT or CIT, the suspect might now look guilty on multiple trials because all the information in the repeated FCT is intermingled with each other. For example, in a crime where only 5 pieces of information are available to the police, 8 out of 20 trials contain stimuli related to the murder weapon. In this way, a knowledgeable innocent suspect would appear guilty on almost half of the trials. This problem becomes even larger if fewer items are being repeated. When there are only 4 pieces of information available, 6 out of 12 trials contain information about 1 item, when there are only 3 pieces of information, 4 out of 6 trials contain information about the same item, etc. I’m curious to hear what you think about this concern.

*Indeed, leakage can have a larger impact, because a single piece of information makes out a larger proportion of the repeated FCT than a traditional FCT, remains intact. We agree that leakage would have a potentially stronger effect on the repeated than the traditional FCT than a traditional FCT, as all repetitions would be compromised.*

*However, the test procedure is adapted to prevent cases from leakage to occur. Aside from the fact that the examiner should take great care to avoid questions about information that was leaked the test also contains another procedure to deal with this. Before the FCT is conducted, the examinee receives all questions and answer alternatives and should indicate to which questions s/he knows the correct answers. These questions are subsequently eliminated from the FCT procedure. Hence, an innocent examinee who has leaked information now has the opportunity to explain how they gained this knowledge and will only encounter questions in the FCT they allege to not know the correct answer to.*

*In experimental research this matter is typically glossed over because the experiment is constructed so that none of the items are affected by leakage. We refer to this procedure with a simple sentence on page 17 lines 1-2:*

*“Before the start of the test, we showed participants all questions with answer alternatives and described the test procedure.”*

*We have modified the section to allude to the leakage problem by changing the section to:*

*“Before the start of the test, we showed participants all questions with answer alternatives and described the test procedure.* ***In practice, this step is used to exclude questions that are affected by leakage of information, for example through news or trial proceedings.”***

*We also updated the FCT & CIT section for study 2 accordingly:*

*“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.for example Denney, 1996).*** *The procedure was as follows:”*

*Finally, we stressed this limitation when we discuss the practical implications in the general discussion by adding the following:*

*“****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.****”*

Reviewer 3

This is an interesting and very well-written manuscript examining the validity of a new method of developing a Forced Choice Test (FCT) in order to detect concealed crime knowledge. I enjoyed reading it. I think both studies are based on carefully planned designs and analyses. The text is both, easy to read and easy to follow.My only suggestions relate to the role of the filler trials that were added in study 2. Adding these item sets actually changes the rationale of the FCT, since the test strategy is based on the idea that respondents try to conceal their knowledge. Therefore, the new FCT in study 1 does not represent the same test as the new FCT in study 2. Perhaps this new FCT test strategy in study 2 works better than the FCT test strategy in study 1 due to the filler items and their effect on answer strategies (higher frequency of strategy 1; comparing Table 1 with Table 3). Therefore, it would be helpful if you elaborated more on the role of the filler items as well as the practical implications of this finding e.g., how to construct adequate answer alternatives that would lead to the same test performance as the FCT you present in your study. As it can be read in the appendix page 33, line 28, filler trials are presented a higher number of times than the critical trials, therefore the quality of those items affect the test performance substantially.

*We agree that filler trials play an important role, but do not think that our design at this point allows us to make recommendations what kind and how many filler trials are best. Instead, we have included the following section in our general discussion to highlight this aspect:*

*We have expanded our discussion on filler trials in study 2:*

*“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****.”*

*We also updated our discussion on filler trials in the general discussion:*

*“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.****”*

Page 21, line 13: the sentence is missing a full stop

*The full stop is present after the “1”. Numbers mentioned in our text do not include dots. I.e. one is written as “1” not “1.”.*

Page 26, line 7: I think it should be “fall” instead of “falls”, since data is plural.

*We have made this change.*

Page 5, line 10 and 11: it would be helpful for the reader, if you mentioned that in the first FCT the proportion of each crime-related answer alternative (e.g., “knife”) remains the same, even though it is paired with multiple non-related answers.

*We agree and clarified the passage as follows:*

*“Correspondingly, traditional pairs from the same category, e.g., “gun” and “knife” or “wallet” and “watch”, are replaced with two seemingly unrelated answers, e.g., “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:”*

*References*

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

Jelicic, M. (2018). Testing claims of crime-related amnesia. *Frontiers in Psychiatry, 9, 617*. <https://doi.org/10.3389/fpsyt.2018.00617>

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>

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

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

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

Orthey, R. (2019). Response strategies of instructed malingerers during forced choice testing: new measures and criteria to detect concealed knowledge and feigned cognitive deficits. *Maastricht:Maastricht University. 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

**Editor First Decision**: Revise & Resubmit

May 20, 2022

Dear Robin Orthey,

I have now read your revised manuscript. I appreciate your careful attention to the concerns the reviewers and I raised. I am happy to provisionally accept your manuscript for submission. However, the reviewers pointed out some small details which I would ask you to adress.

I look forward to receiving your final revision and accepting it for publication in Collabra: Psychology.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at [editorialoffice@collabra.org](mailto:editorialoffice@collabra.org).

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Martin Schultze

**Reviewer 1**

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

I appreciate the additional information on the FCT; I have little knowledge about this method and was not aware of its apparently widespread application.

I have some last small comments below, but I find the manuscript acceptable regardless of how (or whether) these are responded to.

“it has several advantages over the CIT: it is cheap, fast, and easy to use”  
The behavioral “RT-CIT” (with response times alone) cannot be any less cheap, it is hardly less fast, and whether it is less easy to use is debatable. Nonetheless, again, I now understand that the FCT is an established method.

"because performance of innocent participants can be assumed to follow the binomial distribution. "  
This is not self-evident, but I accept that it was shown true by the cited papers.

Lastly, although in case of combining merely two variables it seems permissible to forego proper machine learning procedures, it might be noted that the generalizability of the obtained improvement is uncertain without cross-validation on other data.

**Reviewer 2**

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  |  | ✔ |

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

This is an interesting and very well- written manuscript examining the validity of a new method of developing a Forced Choice Test (FCT) in order to detect concealed crime knowledge. I have been reviewer in the first round. I think that the manuscript substantially improved regarding its scientific quality. I also found all my points being answered satisfactorily. Nevertheless, I have some minor remarks left.

1. Page 5, line 7: What was absolute the number of trials the “minimal test length” refers to?
2. Page 4, line 12. I would be interested in the absolute number of trials that are necessary to conduct the “runs-test” in order to discriminate between fake and true randomness. Is it a score that is usually calculated in the FCT? Just out of curiosity, is it possible to calculate this test with your FCT data too and could it add information to the test score you are investigating in the current study? How is it different to your examination with the binomial distribution?
3. P. 7, line 6: How did you select the 100 items from the possible 380 possible pairs? Did you ensure somehow, that the 100 items represented all 20 pieces of information that you were using for construction?  
   Method
4. P. 6, line 23: You might add a comma or rephrase the sentence “picture of a handgun paired with a picture of a rifle or two pictures of public places” since it reads as if the picture of the handgun was paired with a picture of two public places.
5. P. 8, line 4: It is the same idea like in comment Nr. 3: Did you ensure that the first 20 trials represent all 20 pieces of information? Later in the discussion you write, that there might be a different amount of crime related information in the first 20 items due to randomized presentation. Since you want to investigate the effectiveness of the new trial construction, not the effect of the number of information, more elaboration on the issue might be important for the reader. Maybe you could even provide a statistics assessing the amount of information in all trial sets.
6. Page 9, line 4: You might already add the information “the verbal response was recoded” here. Otherwise it might be unclear why you need to recode the answer.
7. Page 10, line 18: please elaborate, how you proceed with the answer “I don’t know” in the calculation of the FCT score.  
   Discussion
8. Page 14, line 5: Additionally to your explanations for the higher accuracy of the traditional FCT compared to other studies is also the new wording of the trials instruction and the added answer alternative “I don’t know”.  
   Study 2
9. Page 15 line 10: there is a typo “he”
10. Discussion: As far as I understood, you changed several aspects in our application of the FCT that differs from the traditional FCT and needed to be applied in practice if one aims to obtain the same accuracy as you do. There is the camera, the wording of the items (“which item is more related to”, the answer alternative “I don’t know”, and in study 2 the filler trials. It would be very helpful for the reader, if you reflect on the all changes in the discussion in a short sentence.

**Author Response**  
Jun 7, 2022

*Dear Editor and reviewers,*

*Thank you for taking the time to consider our manuscript. Please find our answers to your comments below. Changes in the manuscript are highlighted in bold in this document.*

Reviewer 1

I appreciate the additional information on the FCT; I have little knowledge about this method and was not aware of its apparently widespread application.

I have some last small comments below, but I find the manuscript acceptable regardless of how (or whether) these are responded to.

“it has several advantages over the CIT: it is cheap, fast, and easy to use”  
The behavioral “RT-CIT” (with response times alone) cannot be any less cheap, it is hardly less fast, and whether it is less easy to use is debatable. Nonetheless, again, I now understand that the FCT is an established method.

*One could argue that the Forced Choice Tests can be used with only pencil and paper, and therefore do not require any equipment or programming skills, making it cheaper and easier than even the RT-CIT. But we mainly refer to the Autonomic Nervous System based CIT because that is the only variant used in practice, which does rely on equipment.*

"because performance of innocent participants can be assumed to follow the binomial distribution. "  
This is not self-evident, but I accept that it was shown true by the cited papers.

*We may present this as self-evident in our response letter, but in the manuscript we do elaborate on this when introducing the FCT in the beginning (page 2, lines 13 - 15):*

*“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.”*

*Including empirical support for the core assumption that the test items in the FCTs used here can be considered equally plausible (page 6, lines 14 – 15):*

*“The pairs were checked for equal plausibility in Orthey et al., (2018).”*

*And page 17, lines 7 – 15:*

*“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.”*

Lastly, although in case of combining merely two variables it seems permissible to forego proper machine learning procedures, it might be noted that the generalizability of the obtained improvement is uncertain without cross-validation on other data.

*We agree that with two variables, cross validation will yield limited additional information. In the general discussion we do, however, refer to the need of further replication. i.e. page 25 lines 3+ “If replicated, our findings that a diagnostic FCT can be produced with limited information has several potential implications….”*

Reviewer 2

This is an interesting and very well- written manuscript examining the validity of a new method of developing a Forced Choice Test (FCT) in order to detect concealed crime knowledge. I have been reviewer in the first round. I think that the manuscript substantially improved regarding its scientific quality. I also found all my points being answered satisfactorily. Nevertheless, I have some minor remarks left.

1. Page 5, line 7: What was absolute the number of trials the “minimal test length” refers to?

*We clarified the section to mention the exact test length:*

*The only exception is study 2 from Meijer et al. (2007), who achieved a high detection accuracy with a test length* ***of 12 items.***

1. Page 4, line 12. I would be interested in the absolute number of trials that are necessary to conduct the “runs-test” in order to discriminate between fake and true randomness. Is it a score that is usually calculated in the FCT? Just out of curiosity, is it possible to calculate this test with your FCT data too and could it add information to the test score you are investigating in the current study? How is it different to your examination with the binomial distribution?

*This is an interesting question that we are currently investigating in a separate empirical project. Indeed, t**he runs-test also follows the binomial distribution, because the likelihood of alternating between correct and incorrect answer alternatives is the same as selecting the correct answer when truly guessing. Given that the effect-size of the runs test is much lower than the effect size of the underperformance criterion we expect required test length to be larger (Orthey et al., 2018). The reason for testing this in a separate project is that the current project deviates in many ways from a FCT applied in practice and we would like to test this in a more straightforward way.*

1. P. 7, line 6: How did you select the 100 items from the possible 380 possible pairs? Did you ensure somehow, that the 100 items represented all 20 pieces of information that you were using for construction?

*Each of the 20 pieces of information was repeated 5 times to create the 100 test items. We adjusted the text as follows:*

*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**.*

1. P. 6, line 23: You might add a comma or rephrase the sentence “picture of a handgun paired with a picture of a rifle or two pictures of public places” since it reads as if the picture of the handgun was paired with a picture of two public places.

*We made this change, thank you.*

1. P. 8, line 4: It is the same idea like in comment Nr. 3: Did you ensure that the first 20 trials represent all 20 pieces of information? Later in the discussion you write, that there might be a different amount of crime related information in the first 20 items due to randomized presentation. Since you want to investigate the effectiveness of the new trial construction, not the effect of the number of information, more elaboration on the issue might be important for the reader. Maybe you could even provide a statistics assessing the amount of information in all trial sets.

*No, due to the random presentation of trials it is extremely unlikely that the first 20 trials did only contain unique pieces of information. As we acknowledge in the discussion it is very likely that participants encountered less unique pieces of information in the first 20 trials of the repeated FCT than in the standard FCT and the amount may have varied among participants. To address this issue we computed the detection accuracy for the repeated FCT twice. Once only for the first 20 trials, thus ensuring that the standard and repeated FCT are of equal test lengths, and once for all items so that the standard and repeated FCT contain equal amount of information, although the test length of the repeated FCT is increased by fivefold. As we point out in the discussion (p. 25 lines 16+) our novel FCT disentangles test length, i.e. the number of trials, and information richness, i.e. the number of unique pieces of information. Therefore, a comparison with the standard FCT, in which test length and information richness are always the same, is always imbalanced on either of these two dimensions.*

*We believe that a closer examination of the actual information richness in the first 20 trials will not affect the conclusions of this manuscript. Given the limitations we acknowledge, we only feel confident of concluding from study 1 that our repeated FCT can work and is worth to be examined further. We are not convinced that further nuanced conclusions about the test length – information richness trade off are warranted here.*

*However, we have clarified the discussion to stress the potential impact of random trials:*

*“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 random 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.”*

1. Page 9, line 4: You might already add the information “the verbal response was recoded” here. Otherwise it might be unclear why you need to recode the answer.

*We modified the section. It now reads:*

*Their verbal response was audio-recorded and later* ***recoded as described in the previous section.***

1. Page 10, line 18: please elaborate, how you proceed with the answer “I don’t know” in the calculation of the FCT score.

*The option “I don’t know” was only provided for the memory check at the end of the experiment, not in the actual FCT itself. For the purpose of calculating the memory performance we do not distinguish between incorrect answers or “I don’t know”. We simply use it because it is helpful for us to better understand the quality of the materials we constructed, so we can improve them for the future.*

1. Page 14, line 5: Additionally to your explanations for the higher accuracy of the traditional FCT compared to other studies is also the new wording of the trials instruction and the added answer alternative “I don’t know”.

*Firstly, as pointed out, the option “I don’t know” was not present in the actual FCT. As for the wording of the instructions we modified the discussion as follows:*

*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.***

Study 2

1. Page 15 line 10: there is a typo “he”

*Fixed, thank you.*

1. Discussion: As far as I understood, you changed several aspects in our application of the FCT that differs from the traditional FCT and needed to be applied in practice if one aims to obtain the same accuracy as you do. There is the camera, the wording of the items (“which item is more related to”, the answer alternative “I don’t know”, and in study 2 the filler trials. It would be very helpful for the reader, if you reflect on the all changes in the discussion in a short sentence.

*We have expanded the suggestions for future research to include the switch in terms of questions and the role of distractors:*

***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.***

**Editor Final Decision:** Accept

July 25, 2022

Dear Robin Orthey,

I have now had a chance to read over your manuscript “How to Detect Concealed Crime Knowledge in Situations with little Information using the Forced Choice Test”, along with the letter describing the changes you made. Thank you for your responsiveness to the concerns that the reviewers and I raised. I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point.

You will be receiving separate correspondence regarding any production and technical comments, data deposits, as well as publication charges. We work with the Copyright Clearance Center to process any applicable APC charges. Please note that your APC transaction must be completed before your article gets published.

You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely,  
Martin Schultzes