**Peer Review and Communication History**

**MS Title**: Disentangling the Contributions of Repeating Targets, Distractors, and Stimulus Positions to Practice Benefits in d2-like Tests of Attention

**Author Names**: Peter Wühr and Bianca Wühr

**Submitted:** Aug 24, 2022

**Editor First Decision**: Revise & Resubmit

Oct 10, 2022

Dear Dr. Wühr,

I have received two expert reviews of your manuscript “Disentangling the Contributions of Repeating Targets, Distractors, and Stimulus Positions to Practice Benefits in d2-like Tests of Attention” that you submitted to Collabra: Psychology.

I thank Michael B. Steinborn, who chose to sign his review and Reviewer 2, who wishes to remain anonymous, for their thorough and thoughtful comments. Both reviewers were generally positive in their evaluation of the manuscript and recommend revision. The comments are clear and helpful, so I will not repeat them here in detail.

I believe that the test reliability and the alternative explanation (target-distractor interaction) are the most important issues. Please also provide an explanation for the high drop out rate, and elaborate on the rationale behind the experimental design. Both reviewers point out that the introduction is not as concise as it could be, and I agree with them that there is room for improvement here. Please also discuss the limitations mentioned.

I invite you to resubmit a revision that addresses the concerns in the reviews below. If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript.

Yours sincerely,

Jan Philipp Röer
Associate Editor
Collabra: Psychology

Reviewer 1 (Michael B. Steinborn)

Background

This manuscript is concerned with practice effects in the d2 test of sustained attention, in particular, the effect of changing aspects of the task (i.e., turning items to distractors, etc.), which is presented in three experiments. The results provide useful knowledge on the question of how to construct reliable tests and how to minimise practice effects (which is a prerequisite of reliability). The results are very clear and provide a strong and memorable message to the community, as the authors formulate: “… in a nutshell, the results suggest that target learning makes a strong contribution, distractor learning makes a moderate contribution, and contextual learning makes a negligible contribution to the practice gains that are observed when a d2-like test is repeated…..”.

Evaluation

The manuscript is well written, the research question is clear and the experimental approach is straightforward. The introduction is relatively long but not lengthy, rather, it is informative as both theory and empirical studies are described in great detail, though held concise throughout the manuscript. There is little to complain about this study. My comments are outlined below.

1 Theory

though the authors focus on mechanisms of “selection”, it might briefly be stated that the original measurement intention of the d2 test is to measure “sustained attention”. therefore, the test is often used as a test to measure “ego depletion”, to name one study, see Gropel et al., (doi:10.1177/0146167213516636), but there are many others (see Schumann et al., 2022, doi:10.3389/fpsyg.2022.867978, chap. 4.3 to 4.5, for a theoretical review and summary of empirical findings). So in brief, a popular opinion views the d2 test within an energetic-capacity framework, while the authors put more emphasis on the computational aspect of attention control, which is convincing, I only feel that these positions could be mentioned or, if possible, even related to each other a bit.

2 Test reliability

The present experiments purely focus on the experimental aspect of a psychometric test, examining whether the manipulation of targets, distractors, and combination patterns, have a differential effect on test-retest learning. From a psychometric perspective, it is important to know whether the new test forms show similar reliability as compared to the original d2 or more generally, speeded tests of the bourdon type. I suggest (though I would not strictly demand it) showing some information of test-retest correlations of the test, if possible using a MTMM structure, like in the recent study of Schumann et al. doi:10.3389/fpsyg.2022.946626, which I would like to recommend here. Test-retest correlations in the present case (where test variants are not completely identical) could be viewed as the reliability of parallel versions A and B.

3 Line comments

p. 2, intro to the d2 test “the d2 test as visual search process”---------comment: i would agree, though the crucial aspect different from trial-based visual display search is that the participants work in a self-paced way through the list of items, thus the process could be described, in a nutshell, as a visual-scan-and-check process, similar to the process studied in the domain of proofreading.
p. 2, last paragraph “test exists in several languages…”------------comment: it might be specified that the test is a “language free” test and thus considered culture fair, the authors mean that the test handbook is translated into different languages.
p. 3, line “after practice, the test measures diligence instead of ability” ---------comment: after practice, the test measures “skill” in that test, likely not diligence
p. 4, line 9 purity assumptions ----------comments: purity includes two aspects;
Donders’s purity assumption (see Ulrich, Mattes, & Miller, 1999) in self-paced tests, it is more difficult to determine the locus of practice, because of the self-paced working style, and also any manipulation of stages would affect the error rate differently which makes comparison relatively complicated.
Purity doctrine in the psychometric assessment domain assessment researcher often divide tests into classes measuring distinct psychological concepts like “selective attention”, “sustained attention”, “working memory”, “executive attention”, and so on. A empirical correlation between a working-memory test and an attention test would be seen as a lack of discriminance (cf. Schumann et al. doi:10.3389/fpsyg.2022.946626, see Table 1). To name one example, a study of da Silva-Bauer et al. 2022 showed that the d2 test correlates with working memory measures which according to the authors, is a violation of the purity of measurement intention. However, the d2 test was originally intended to measure elementary cognitive ability as broadband construct and the term attention was not meant as a very distinct or discriminate concept.
p. 5, line 7 d2 test is a visual search task----------comment: visual-check-and-scan
p. 13, line 4 time intervals between tasks: ----------comment: not only the time intervals but the conception as self-paced test is crucial and that the participants can see adjacent items at once.
p. 14, sample dropout experiment 1, 159 to 142 to 133 (after Tukey test) shows a relatively high dropout rate, the same holds for experiment 2 too---------comment: I am okay with the authors argument, this is only a personal note: this is a very high dropout rate. It might be stated that criteria in assessment research are different from cognitive basic research. Hagemeister argued this way: if you have participants with high error rate in a cognitive experiment, then you can say, okay, my model does not work with that high error rate, so i have to exclude the participants. In an assessment context, the error score is a defined score index to measure an aspect of performance and participants with high error score would simply score worse on that performance dimension. in the assessment context, particpants would only be excluded if the show a clear instruction deviant behavior but not because they are slow or error prone.
p. 18, outliers via Tukey test? ----------comment: i would exclude participants only if they exhibit a pattern that indicates instruction-deviant behavior, to exclude participants with low or high score might reduce reliability as one could argue that they naturally occur in the population.
p. 25, figures 1 and 2 and … if possible, the figures could be presented in a more economic fashion, for example, together using panels A, B, C. etc. this would give overview and the results could be directly compared.
p. 32, results in a nutshell In a nutshell, the results suggest that target learning makes a strong contribution, distractor learning makes a moderate contribution, and contextual learning makes a negligible contribution to the practice gains that are observed when a d2-like test is repeated.
p. 35-36, conclusion i would end the manuscript with the essence in a “nutshell”, which was said earlier, but it should be said here as a take home message.

Reviewer 2

The authors are reporting three experiments on different types practice effects in the d2 test of selective attention. Specifically, they investigate to what extent the repetition of targets (e.g., any “d” with two dashes), the repetition of distractors (e.g., “d” with one or three dashes, or “p”, “b”, “q” with two dashes), and repetition of stimulus configurations / positions contribute to practice effects on hit rates, misses and false alarms in the d2 test. The manuscript is definitely well written and very clear, and the results seem highly relevant in particular for practical purposes such as the development and improvement of classical “pen and paper” tests of selective attention. I believe that this work should be suitable for publication in Collabra: Psychology, and it might be interesting to a broad range of readers. However, I also have a few comments and suggestions on how to further improve the paper.

The introduction is quite long, containing several iterations of similar ideas. I think the readability would benefit from a shorter and more concise introduction. For instance, the authors refer to the training study by Fisk et al. (1991) multiple times (on pp. 8-10) using very similar wording, and it might be difficult for the readers to extract the differences (i.e., target-repetition vs. distractor-repetition conditions). Maybe it would be better to first explain the different experimental conditions from Fisk et al. (1991) together in one section, and then discuss the implications for learning about targets vs. distractors with regard to the different theoretical models?

I also think that other parts of the introduction could be elaborated further: On p. 4, the authors may want to present a bit more information on the “representative tasks” that were used to measure the three processes (perceptual speed, simple mental operation, motor speed), and why the tasks cannot be considered “process-pure”. Moreover, I would like to see how these three processes relate to the components of the d2 that were investigated in this paper (targets, distractors, stimulus configurations), which are introduced in the second paragraph on p. 4. Speaking of this paragraph: I think the authors should also add some further information on the exact targets that were used in session 1 and 2. The degree of practice effects in the target-change condition (i.e., transfer effects) might certainly depend on the similarity of the targets in the two sessions. I think this could be discussed a bit more (e.g., referring to findings from the perceptual learning literature).

More comments on the introduction: On p. 5 (bottom), the authors could mention whether (and why) the d2 can be considered a feature-search task as well. On p. 12, it should be explained whether stimulus learning (from the study by Le Dantec et al., 2012) refers to target or distractor practice effects, or whether it can’t be disentangled. I the latter case, the results of that study may only partially justify the approach to investigate the effects of targets, distractors, and stimulus configurations separately (i.e., to not investigate possible interactions between target and distractor processing).

The rationale behind the particular experimental design might seem a bit counterintuitive. The aim of Experiment 1 is to investigate the contribution of target repetitions, and the authors are contrasting a “target-repetition condition” with a “target-change condition”. However, I think the first condition is actually a “repeat everything condition” that is contrasted with a “repeat everything but the targets” condition. I was wondering the the authors decided not to use an alternative design in which for one group only the targets are repeated (and everything else is changed from session 1 to 2), whereas for the other (control) group everything is changed? I think this would be an even purer test of the contribution of target repetitions to practice effects. Maybe the authors decided to go with the reversed design because the practical goal is to identify conditions with minimized practice effects, but (if this is the case) this could be made more explicit. Related to this point, I wonder whether the first sentence on p. 14 is true. Couldn’t it be the case that practice effects result from a repetition of the whole configuration (of targets x distractors x positions) in the “target-repetition condition” rather than from target repetitions alone? In other words, I don’t think it cannot be ruled out entirely that the improvements are due to possible interactions of targets with distractors (or positions) in the “target-repetition condition”. This limitation could be discussed.

Given that the two groups of participants were tested together in large groups and that they were (all?) told about the changes for some participants (p. 16), I wonder whether demand characteristics could have played a role. The authors could discuss this in the General Discussion as I believe it applied to all three experiments.

Tables 1-3: I was not sure what the numbers mean (probably absolute numbers of hits/KL, false alarms, and misses?), but I strongly suggest to present the data as percentages (maybe in addition to the absolute numbers) in order to allow for easier comparisons across experiments (the tasks in the three experiments differ slighly in the maximum number of targers and distractors).

Figures 1-3: Error bars should be added (e.g., showing standard errors of the mean).

It seems that a pilot study was conducted to avoid ceiling effects in Exp. 2 (p. 23), but not for the tasks in the other two experiments. The authors may want to add an explanation for this. I addume, maybe the tasks had already been used in other studies? Also, there seems to be additional variation from session 1 to session 2 in Exp. 2 (with different spatial arrangements of the stimuli in the versions A1/B1 vs A2/B2), which was not there in Exp. 1 and 3. Why was this additional variation introduced, and could it account for some of the descrepancies in results?

Looking at the specific stimuli that were used as targets and distractors in the different experiments, I wonder whether the similarity between session 1 and 2 (in the change condition) was higher in Exp. 2 than in Exp. 1. That is, the tartget set changed from [“d d” ‘d’] to [“p p” ‘p’] in Exp. 1 (i.e., a completely new target letter), whereas the distractor set changed from [‘d "d’ ‘p’ “p p”] to [d’ "d’ ‘q’ “q q”] in Exp. 2 (i.e., quite some overlap in distractor characters). So could it be the case that the smaller difference between the two conditions in terms of practice effects was (partially) due to differences in stimulus similarity? I think the issue of similarity and generalization deserves some discussion. For instance, on p. 33, the authors state that “repeating distractors seems to play a minor role”. I am not entirely convinced by this statement as the similarity of the two sessions in Exp. 2 may have been higher than in Exp. 1, and this could at least have contributed to the differences in practice effects between Exp. 1 and 2. The authors may want to discuss this a bit.

Minor points

p. 9, “Hence, its unclear whether” -> “Hence, it is unclear whether”

p. 20, bottom, “KL” comes a bit surprising. It has been introduced early in the paper (p. 2), but not used in the more recent results section. I think the authors should use only one term (hits or KL) consistently throughout the manuscript.

p. 25, looking at Fig. 2, I wondered whether the pre-test difference in Hits was significant? The same is true for the difference in false alarms and misses during session 1 (Table 2).

p. 28, line 4: Maybe I am wrong, but I think the three targets should be: “d” with two dashes above, “d” with two dashes below, and a “d” with one dash above and one below. Otherwise there seems to be overlap with the distractors.

p. 28 middle paragraph: Why was one row repeated more often than the other two rows? Couldn’t the authors just have used 15 rows (5 repetitions of each row)? As it currently is, participants practiced the first row more than the second and third row. Is there any indication in the data of a stronger performance benefit (in the position repetition condition) for the first row?

The authors often refer to “numerical” differences, as opposed to “significant” differences (e.g., top of p. 26). I think this may be misleading in some places, in particular in the discussion section (such as in the discussion of Exp. 3, p. 31): I don’t think it is important to discuss a small difference in improvements when it is not significant (it would be quite unlikely to not observe a small “numerical difference” with finite samples).

p. 32, third line of second paragraph: “this learning improves performance” (s missing after “improve”)

p. 36: Attentional improvements and perceptual learning could also be mentioned as possible variables contributing to stimulus-independent learning.

# 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.

Review of Manuscript; R0; Disentangling the Contributions of Repeating Targets, Distractors, and Stimulus Positions to Practice Benefits in d2-like Tests of Attention.
Journal: collabra: psychology
Reviewer: Michael B. Steinborn

Background:

This manuscript is concerned with practice effects in the d2 test of sustained attention, in particular, the effect of changing aspects of the task (i.e., turning items to distractors, etc.), which is presented in three experiments. The results provide useful knowledge on the question of how to construct reliable tests and how to minimise practice effects (which is a prerequisite of reliability). The results are very clear and provide a strong and memorable message to the community, as the authors formulate: “… in a nutshell, the results suggest that target learning makes a strong contribution, distractor learning makes a moderate contribution, and contextual learning makes a negligible contribution to the practice gains that are observed when a d2-like test is repeated…..”.

Evaluation:

The manuscript is well written, the research question is clear and the experimental approach is straightforward. The introduction is relatively long but not lengthy, rather, it is informative as both theory and empirical studies are described in great detail, though held concise throughout the manuscript. There is little to complain about this study. My comments are outlined below.

(-1-) theory

though the authors focus on mechanisms of “selection”, it might briefly be stated that the original measurement intention of the d2 test is to measure “sustained attention”. therefore, the test is often used as a test to measure “ego depletion”, to name one study, see Gropel et al., (doi:10.1177/0146167213516636), but there are many others (see Schumann et al., 2022, doi:10.3389/fpsyg.2022.867978, chap. 4.3 to 4.5, for a theoretical review and summary of empirical findings). So in brief, a popular opinion views the d2 test within an energetic-capacity framework, while the authors put more emphasis on the computational aspect of attention control, which is convincing, I only feel that these positions could be mentioned or, if possible, even related to each other a bit.

(-2-) test reliability

The present experiments purely focus on the experimental aspect of a psychometric test, examining whether the manipulation of targets, distractors, and combination patterns, have a differential effect on test-retest learning. From a psychometric perspective, it is important to know whether the new test forms show similar reliability as compared to the original d2 or more generally, speeded tests of the bourdon type. I suggest (though I would not strictly demand it) showing some information of test-retest correlations of the test, if possible using a MTMM structure, like in the recent study of Schumann et al. doi:10.3389/fpsyg.2022.946626, which I would like to recommend here. Test-retest correlations in the present case (where test variants are not completely identical) could be viewed as the reliability of parallel versions A and B.

(-3-) line comments

p. 2, intro to the d2 test
“the d2 test as visual search process”---------comment: i would agree, though the crucial aspect different from trial-based visual display search is that the participants work in a self-paced way through the list of items, thus the process could be described, in a nutshell, as a visual-scan-and-check process, similar to the process studied in the domain of proofreading.

p. 2, last paragraph
“test exists in several languages…”------------comment: it might be specified that the test is a “language free” test and thus considered culture fair, the authors mean that the test handbook is translated into different languages.

p. 3, line
“after practice, the test measures diligence instead of ability” ---------comment: after practice, the test measures “skill” in that test, likely not diligence

p. 4, line 9
purity assumptions ----------comments: purity includes two aspects;

Donders’s purity assumption (see Ulrich, Mattes, & Miller, 1999)
in self-paced tests, it is more difficult to determine the locus of practice, because of the self-paced working style, and also any manipulation of stages would affect the error rate differently which makes comparison relatively complicated.

Purity doctrine in the psychometric assessment domain
assessment researcher often divide tests into classes measuring distinct psychological concepts like “selective attention”, “sustained attention”, “working memory”, “executive attention”, and so on. A empirical correlation between a working-memory test and an attention test would be seen as a lack of discriminance (cf. Schumann et al. doi:10.3389/fpsyg.2022.946626, see Table 1). To name one example, a study of da Silva-Bauer et al. 2022 showed that the d2 test correlates with working memory measures which according to the authors, is a violation of the purity of measurement intention. However, the d2 test was originally intended to measure elementary cognitive ability as broadband construct and the term attention was not meant as a very distinct or discriminate concept.

p. 5, line 7
d2 test is a visual search task----------comment: visual-check-and-scan

p. 13, line 4
time intervals between tasks: ----------comment: not only the time intervals but the conception as self-paced test is crucial and that the participants can see adjacent items at once.

p. 14, sample dropout
experiment 1, 159 to 142 to 133 (after Tukey test) shows a relatively high dropout rate, the same holds for experiment 2 too---------comment: I am okay with the authors argument, this is only a personal note: this is a very high dropout rate. It might be stated that criteria in assessment research are different from cognitive basic research. Hagemeister argued this way: if you have participants with high error rate in a cognitive experiment, then you can say, okay, my model does not work with that high error rate, so i have to exclude the participants. In an assessment context, the error score is a defined score index to measure an aspect of performance and participants with high error score would simply score worse on that performance dimension. in the assessment context, particpants would only be excluded if the show a clear instruction deviant behavior but not because they are slow or error prone.

p. 18, outliers via Tukey test?
----------comment: i would exclude participants only if they exhibit a pattern that indicates instruction-deviant behavior, to exclude participants with low or high score might reduce reliability as one could argue that they naturally occur in the population.

p. 25, figures 1 and 2 and …
if possible, the figures could be presented in a more economic fashion, for example, together using panels A, B, C. etc. this would give overview and the results could be directly compared.

p. 32, results in a nutshell
In a nutshell, the results suggest that target learning makes a strong contribution, distractor learning makes a moderate contribution, and contextual learning makes a negligible contribution to the practice gains that are observed when a d2-like test is repeated.

p. 35-36, conclusion
i would end the manuscript with the essence in a “nutshell”, which was said earlier, but it should be said here as a take home message.

# 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 authors are reporting three experiments on different types practice effects in the d2 test of selective attention. Specifically, they investigate to what extent the repetition of targets (e.g., any “d” with two dashes), the repetition of distractors (e.g., “d” with one or three dashes, or “p”, “b”, “q” with two dashes), and repetition of stimulus configurations / positions contribute to practice effects on hit rates, misses and false alarms in the d2 test. The manuscript is definitely well written and very clear, and the results seem highly relevant in particular for practical purposes such as the development and improvement of classical “pen and paper” tests of selective attention. I believe that this work should be suitable for publication in Collabra: Psychology, and it might be interesting to a broad range of readers. However, I also have a few comments and suggestions on how to further improve the paper.

* The introduction is quite long, containing several iterations of similar ideas. I think the readability would benefit from a shorter and more concise introduction. For instance, the authors refer to the training study by Fisk et al. (1991) multiple times (on pp. 8-10) using very similar wording, and it might be difficult for the readers to extract the differences (i.e., target-repetition vs. distractor-repetition conditions). Maybe it would be better to first explain the different experimental conditions from Fisk et al. (1991) together in one section, and then discuss the implications for learning about targets vs. distractors with regard to the different theoretical models?
* I also think that other parts of the introduction could be elaborated further: On p. 4, the authors may want to present a bit more information on the “representative tasks” that were used to measure the three processes (perceptual speed, simple mental operation, motor speed), and why the tasks cannot be considered “process-pure”. Moreover, I would like to see how these three processes relate to the components of the d2 that were investigated in this paper (targets, distractors, stimulus configurations), which are introduced in the second paragraph on p. 4. Speaking of this paragraph: I think the authors should also add some further information on the exact targets that were used in session 1 and 2. The degree of practice effects in the target-change condition (i.e., transfer effects) might certainly depend on the similarity of the targets in the two sessions. I think this could be discussed a bit more (e.g., referring to findings from the perceptual learning literature).
* More comments on the introduction: On p. 5 (bottom), the authors could mention whether (and why) the d2 can be considered a feature-search task as well. On p. 12, it should be explained whether stimulus learning (from the study by Le Dantec et al., 2012) refers to target or distractor practice effects, or whether it can’t be disentangled. I the latter case, the results of that study may only partially justify the approach to investigate the effects of targets, distractors, and stimulus configurations separately (i.e., to not investigate possible interactions between target and distractor processing).
* The rationale behind the particular experimental design might seem a bit counterintuitive. The aim of Experiment 1 is to investigate the contribution of target repetitions, and the authors are contrasting a “target-repetition condition” with a “target-change condition”. However, I think the first condition is actually a “repeat everything condition” that is contrasted with a “repeat everything but the targets” condition. I was wondering the the authors decided not to use an alternative design in which for one group only the targets are repeated (and everything else is changed from session 1 to 2), whereas for the other (control) group everything is changed? I think this would be an even purer test of the contribution of target repetitions to practice effects. Maybe the authors decided to go with the reversed design because the practical goal is to identify conditions with minimized practice effects, but (if this is the case) this could be made more explicit. Related to this point, I wonder whether the first sentence on p. 14 is true. Couldn’t it be the case that practice effects result from a repetition of the whole configuration (of targets x distractors x positions) in the “target-repetition condition” rather than from target repetitions alone? In other words, I don’t think it cannot be ruled out entirely that the improvements are due to possible interactions of targets with distractors (or positions) in the “target-repetition condition”. This limitation could be discussed.
* Given that the two groups of participants were tested together in large groups and that they were (all?) told about the changes for some participants (p. 16), I wonder whether demand characteristics could have played a role. The authors could discuss this in the General Discussion as I believe it applied to all three experiments.
* Tables 1-3: I was not sure what the numbers mean (probably absolute numbers of hits/KL, false alarms, and misses?), but I strongly suggest to present the data as percentages (maybe in addition to the absolute numbers) in order to allow for easier comparisons across experiments (the tasks in the three experiments differ slighly in the maximum number of targers and distractors).
* Figures 1-3: Error bars should be added (e.g., showing standard errors of the mean).
* It seems that a pilot study was conducted to avoid ceiling effects in Exp. 2 (p. 23), but not for the tasks in the other two experiments. The authors may want to add an explanation for this. I addume, maybe the tasks had already been used in other studies? Also, there seems to be additional variation from session 1 to session 2 in Exp. 2 (with different spatial arrangements of the stimuli in the versions A1/B1 vs A2/B2), which was not there in Exp. 1 and 3. Why was this additional variation introduced, and could it account for some of the descrepancies in results?
* Looking at the specific stimuli that were used as targets and distractors in the different experiments, I wonder whether the similarity between session 1 and 2 (in the change condition) was higher in Exp. 2 than in Exp. 1. That is, the tartget set changed from [“d d” ‘d’] to [“p p” ‘p’] in Exp. 1 (i.e., a completely new target letter), whereas the distractor set changed from [‘d "d’ ‘p’ “p p”] to [d’ "d’ ‘q’ “q q”] in Exp. 2 (i.e., quite some overlap in distractor characters). So could it be the case that the smaller difference between the two conditions in terms of practice effects was (partially) due to differences in stimulus similarity? I think the issue of similarity and generalization deserves some discussion. For instance, on p. 33, the authors state that “repeating distractors seems to play a minor role”. I am not entirely convinced by this statement as the similarity of the two sessions in Exp. 2 may have been higher than in Exp. 1, and this could at least have contributed to the differences in practice effects between Exp. 1 and 2. The authors may want to discuss this a bit.

Minor points

* p. 9, “Hence, its unclear whether” -> “Hence, it is unclear whether”
* p. 20, bottom, “KL” comes a bit surprising. It has been introduced early in the paper (p. 2), but not used in the more recent results section. I think the authors should use only one term (hits or KL) consistently throughout the manuscript.
* p. 25, looking at Fig. 2, I wondered whether the pre-test difference in Hits was significant? The same is true for the difference in false alarms and misses during session 1 (Table 2).
* p. 28, line 4: Maybe I am wrong, but I think the three targets should be: “d” with two dashes above, “d” with two dashes below, and a “d” with one dash above and one below. Otherwise there seems to be overlap with the distractors.
* p. 28 middle paragraph: Why was one row repeated more often than the other two rows? Couldn’t the authors just have used 15 rows (5 repetitions of each row)? As it currently is, participants practiced the first row more than the second and third row. Is there any indication in the data of a stronger performance benefit (in the position repetition condition) for the first row?
* The authors often refer to “numerical” differences, as opposed to “significant” differences (e.g., top of p. 26). I think this may be misleading in some places, in particular in the discussion section (such as in the discussion of Exp. 3, p. 31): I don’t think it is important to discuss a small difference in improvements when it is not significant (it would be quite unlikely to not observe a small “numerical difference” with finite samples).
* p. 32, third line of second paragraph: “this learning improves performance” (s missing after “improve”)
* p. 36: Attentional improvements and perceptual learning could also be mentioned as possible variables contributing to stimulus-independent learning.

**Author Response**
Jan 7, 2023

**Responses to Reviewer comments**

(Revision of “Disentangling the contributions of repeating targets, distractors, and stimulus positions to practice benefits in d2-like tests of attention” for Collabra: Psychology)

**Responses to Reviewer 1**

The manuscript is well written, the research question is clear and the experimental approach is straightforward. The introduction is relatively long but not lengthy, rather, it is informative as both theory and empirical studies are described in great detail, though held concise throughout the manuscript. There is little to complain about this study. My comments are outlined below.

Author response: We thank Reviewer 1 for his positive evaluation of our work, and for his constructive comments on the previous version of our manuscript that helped us improving our paper. Changes to the manuscript have been marked by red font color.

(1) Theory. Though the authors focus on mechanisms of “selection”, it might briefly be stated that the original measurement intention of the d2 test is to measure “sustained attention”. therefore, the test is often used as a test to measure “ego depletion”, to name one study, see Gropel et al., (doi:10.1177/0146167213516636), but there are many others (see Schumann et al., 2022, doi:10.3389/fpsyg.2022.867978, chap. 4.3 to 4.5, for a theoretical review and summary of empirical findings). So, in brief, a popular opinion views the d2 test within an energetic-capacity framework, while the authors put more emphasis on the computational aspect of attention control, which is convincing, I only feel that these positions could be mentioned or, if possible, even related to each other a bit.

Author response: We thank the reviewer for referring to the literature on “ego depletion”, and to the fact that letter-cancellation tests have been used for inducing or measuring “ego depletion”. However, when inspecting that literature, we learned that the existence of the “ego depletion” effect has been empirically challenged (e.g., Carter et al., 2015; Etherton et al., 2018; Hagger et al., 2016; Vohs et al., 2021), and that the use of letter-cancellation tests for inducing or measuring “ego deplection”, or “cognitive capacity”, has also been called into question (e.g., Wimmer, Stirk, & Hancock, 2017; Wimmer, Come, Hancock, & Wennekers, 2019). Due to the unclear empirical support of both the “ego depletion” effect per se, and the use of letter-cancellation tests for measuring “ego depletion” or cognitive capacity, we decided not to mention this literature in our manuscript. A further reason for our decision is that introducing the capacity view of attention is neither necessary for motivating our experiments, nor for discussing our results.

(2) Test reliability. The present experiments purely focus on the experimental aspect of a psychometric test, examining whether the manipulation of targets, distractors, and combination patterns, have a differential effect on test-retest learning. From a psychometric perspective, it is important to know whether the new test forms show similar reliability as compared to the original d2 or more generally, speeded tests of the bourdon type. I suggest (though I would not strictly demand it) showing some information of test-retest correlations of the test, if possible, using a MTMM structure, like in the recent study of Schumann et al. doi:10.3389/fpsyg.2022.946626, which I would like to recommend here. Test-retest correlations in the present case (where test variants are not completely identical) could be viewed as the reliability of parallel versions A and B.

Author response: We agree that presenting information about the test-retest correlations observed in our experiments might be informative to some readers. Therefore, we added an appendix to the revised manuscript where we present both correlations between the same measure in different sessions (test-retest reliabilities) and correlations between different measures in the same session. As suggested by Reviewer 1, these tables have the MTMM structure as used in Steinborn et al. (2018) and in Schumann et al. (2022). We do not further discuss these correlations because test-retest reliabilities are not the focus of our paper, and we did not want further increase the length of our manuscript.

(3) Line comments

(3.1) p. 2, intro to the d2 test: “the d2 test as visual search process”---------comment: i would agree, though the crucial aspect different from trial-based visual display search is that the participants work in a self-paced way through the list of items, thus the process could be described, in a nutshell, as a visual-scan-and-check process, similar to the process studied in the domain of proofreading.

Author response: We tend to disagree on this point. In a single trial of a typical visual-search experiment (e.g., Treisman & Gelade, 1980), participants are shown displays containing multiple stimuli, and a target may be presented at an unknown location between several distractors. Participants then work in a self-paced way through each display until a target is found or not. Hence, for each trial, search is self-paced in visual-search tasks as well. The difference is that the trial stops, and a new display is presented, when a target was found or when all items were identified as distractors.

(3.2) p. 2, last paragraph: “test exists in several languages…”------------comment: it might be specified that the test is a “language free” test and thus considered culture fair, the authors mean that the test handbook is translated into different languages.

Author response: We have modified the text accordingly.

(3.3) p. 3, line: “after practice, the test measures diligence instead of ability” ---------comment: after practice, the test measures “skill” in that test, likely not diligence.

Author response: We see the point, and therefore we deleted this sentence.

(3.4) p. 4, line 9: purity assumptions ----------comments: purity includes two aspects; (a) Donders’s purity assumption (see Ulrich, Mattes, & Miller, 1999): in self-paced tests, it is more difficult to determine the locus of practice, because of the self-paced working style, and also any manipulation of stages would affect the error rate differently which makes comparison relatively complicated. (b) Purity doctrine in the psychometric assessment domain: assessment researcher often divide tests into classes measuring distinct psychological concepts like “selective attention”, “sustained attention”, “working memory”, “executive attention”, and so on. An empirical correlation between a working-memory test and an attention test would be seen as a lack of discriminance (cf. Schumann et al. doi:10.3389/fpsyg.2022.946626, see Table 1). To name one example, a study of da Silva-Bauer et al. 2022 showed that the d2 test correlates with working memory measures which according to the authors, is a violation of the purity of measurement intention. However, the d2 test was originally intended to measure elementary cognitive ability as broadband construct and the term attention was not meant as a very distinct or discriminate concept.

Author response: We added a statement explaining why the tasks used by Blotenberg and Schmidt-Atzert (2019) were not process pure.

(3.5) p. 5, line 7: d2 test is a visual search task----------comment: visual-check-and-scan.

Author response: We would like to stick to the well-established term “visual-search task” (or “visual search task”). This term is well established in the literature about visual attention (see, e.g., Cave & Bichot, 1999; Chan & Hayward, 2013; for reviews), and the d2 test is a prototypical example of a visual-search task, when viewed from the perspective of an experimental psychologist. In contrast, the term “visual-check-and-scan” task is not familiar to us, which may be related to the fact that this term is probably more frequent in the assessment literature. In this article, however, we wish to analyze performance in the d2 test from the viewpoint of the experimental psychologist, and therefore wish to use terms that are common and well-established in the experimental attention literature.

(3.6) p. 13, line 4: time intervals between tasks: ----------comment: not only the time intervals but the conception as self-paced test is crucial and that the participants can see adjacent items at once.

Author response: We agree, and added a corresponding sentence to the manuscript.

(3.7) p. 14, sample dropout: experiment 1, 159 to 142 to 133 (after Tukey test) shows a relatively high dropout rate, the same holds for experiment 2 too---------comment: I am okay with the authors argument, this is only a personal note: this is a very high dropout rate. It might be stated that criteria in assessment research are different from cognitive basic research. Hagemeister argued this way: if you have participants with high error rate in a cognitive experiment, then you can say, okay, my model does not work with that high error rate, so i have to exclude the participants. In an assessment context, the error score is a defined score index to measure an aspect of performance and participants with high error score would simply score worse on that performance dimension. in the assessment context, participants would only be excluded if the show a clear instruction deviant behavior but not because they are slow or error prone.

Author response: We agree that Experiments 1 and 2 show relatively high dropout rates. However, in both experiments, the larger part of the dropout resulted from participants who did not show up for the second session. This could not be avoided since participation in our experiments was, of course, voluntary. We then excluded 7 participants (5%) from Experiment 1, and four participants (4%) in Experiment 2, because of outliers in performance, and these numbers seem moderate when compared to the experimental literature. In Experiment 1, three of the excluded participants achieved less than 50 hits, and four participants made more than 80 errors (either false alarms or misses). In our view, these numbers reflect instruction-deviant behavior because it seems unlikely that (motivated) students would perform that bad under normal conditions. We added short sentences describing the excluded datasets for all experiments.

(3.8) p. 18, outliers via Tukey test: ----------comment: i would exclude participants only if they exhibit a pattern that indicates instruction-deviant behavior, to exclude participants with low or high score might reduce reliability as one could argue that they naturally occur in the population.

Author response: Please see our response to comment 3.7.

(3.9) p. 25, figures 1 and 2 and …: if possible, the figures could be presented in a more economic fashion, for example, together using panels A, B, C. etc. this would give overview and the results could be directly compared.

Author response: We combined previous Figures 1-3 into a single Figure 1 with three panels (A-C).

(3.10) p. 32, results in a nutshell: In a nutshell, the results suggest that target learning makes a strong contribution, distractor learning makes a moderate contribution, and contextual learning makes a negligible contribution to the practice gains that are observed when a d2-like test is repeated.

Author response: Please see our response to comment 3.11.

(3.11) p. 35-36, conclusion: I would end the manuscript with the essence in a “nutshell”, which was said earlier, but it should be said here as a take home message.

Author response: We agree, and we moved the summary to the end of the manuscript, where we added a new “conclusion” section.

**Responses to Reviewer 2**

The authors are reporting three experiments on different types practice effects in the d2 test of selective attention. Specifically, they investigate to what extent the repetition of targets (e.g., any “d” with two dashes), the repetition of distractors (e.g., “d” with one or three dashes, or “p”, “b”, “q” with two dashes), and repetition of stimulus configurations / positions contribute to practice effects on hit rates, misses and false alarms in the d2 test. The manuscript is definitely well written and very clear, and the results seem highly relevant in particular for practical purposes such as the development and improvement of classical “pen and paper” tests of selective attention. I believe that this work should be suitable for publication in Collabra: Psychology, and it might be interesting to a broad range of readers. However, I also have a few comments and suggestions on how to further improve the paper.

Author response: We thank the author for his evaluation of our research, and for insightful comments on the previous manuscript that helped us improving the paper. Changes to the manuscript have been marked by red font color.

1. The introduction is quite long, containing several iterations of similar ideas. I think the readability would benefit from a shorter and more concise introduction. For instance, the authors refer to the training study by Fisk et al. (1991) multiple times (on pp. 8-10) using very similar wording, and it might be difficult for the readers to extract the differences (i.e., target-repetition vs. distractor-repetition conditions). Maybe it would be better to first explain the different experimental conditions from Fisk et al. (1991) together in one section, and then discuss the implications for learning about targets vs. distractors with regard to the different theoretical models?

Author response: It is true that the introduction of our manuscript is quite long. The main reason for the long introduction is our wish to provide a comprehensive review of (a) short-term effects of repeating stimuli and stimulus positions on performance in simple filtering tasks, and of (b) long-term effects of repeating stimuli and stimulus configurations in visual-search tasks. In doing so we wanted to point out similarities in empirical findings and in theoretical approaches. It is also true that the study of Fisk et al. (1991) is mentioned twice, albeit different conditions are discussed in different places. Our main motivation for doing that is our wish to structure the introduction in a way that matches the experiments. We believe this structure of the introduction provides a clear theoretical background for our experiments, and therefore we do not wish to change this structure. Nevertheless, we carefully reviewed the introduction and streamlined the text as much as possible.

1. I also think that other parts of the introduction could be elaborated further: On p. 4, the authors may want to present a bit more information on the “representative tasks” that were used to measure the three processes (perceptual speed, simple mental operation, motor speed), and why the tasks cannot be considered “process-pure”. Moreover, I would like to see how these three processes relate to the components of the d2 that were investigated in this paper (targets, distractors, stimulus configurations), which are introduced in the second paragraph on p. 4. Speaking of this paragraph: I think the authors should also add some further information on the exact targets that were used in session 1 and 2. The degree of practice effects in the target-change condition (i.e., transfer effects) might certainly depend on the similarity of the targets in the two sessions. I think this could be discussed a bit more (e.g., referring to findings from the perceptual learning literature).

Author response: (a) In the revised version of the manuscript, we present more information on the “representative tasks” that were used to measure the three processes in the study of Blotenberg and Schmidt-Atzert (2019). (b) We added a statement explaining why the tasks cannot be considered “process-pure.” (c) We now explain how the three processes, investigated by Blotenberg and Schmidt-Atzert (2019) relate to the components in the d2 that were investigated in our study. (d) We added some further information on the exact targets that were used in session 1 and 2 of the Wühr (2019) study. (e) We now discuss the issue of stimulus similarity in a new paragraph, called “methodological issues,” in the GD.

1. More comments on the introduction: On p. 5 (bottom), the authors could mention whether (and why) the d2 can be considered a feature-search task as well. On p. 12, it should be explained whether stimulus learning (from the study by Le Dantec et al., 2012) refers to target or distractor practice effects, or whether it can’t be disentangled. I the latter case, the results of that study may only partially justify the approach to investigate the effects of targets, distractors, and stimulus configurations separately (i.e., to not investigate possible interactions between target and distractor processing).

Author response: We added two sentences explaining why the d2 test requires conjunction search, but cannot be mastered by using feature search. Moreover, concerning the study of Le Dantec (2012), we added two sentences explaining how the authors disentangled target and distractor learning in test performance.

1. The rationale behind the particular experimental design might seem a bit counterintuitive. The aim of Experiment 1 is to investigate the contribution of target repetitions, and the authors are contrasting a “target-repetition condition” with a “target-change condition”. However, I think the first condition is actually a “repeat everything condition” that is contrasted with a “repeat everything but the targets” condition. I was wondering why the authors decided not to use an alternative design in which for one group only the targets are repeated (and everything else is changed from session 1 to 2), whereas for the other (control) group everything is changed? I think this would be an even purer test of the contribution of target repetitions to practice effects. Maybe the authors decided to go with the reversed design because the practical goal is to identify conditions with minimized practice effects, but (if this is the case) this could be made more explicit. Related to this point, I wonder whether the first sentence on p. 14 is true. Couldn’t it be the case that practice effects result from a repetition of the whole configuration (of targets x distractors x positions) in the “target-repetition condition” rather than from target repetitions alone? In other words, I don’t think it cannot be ruled out entirely that the improvements are due to possible interactions of targets with distractors (or positions) in the “target-repetition condition”. This limitation could be discussed.

Author response: (a) Concerning the rationale behind our experimental design, we agree that comparing a “change everything” condition with a “change everything but the targets” condition would also provide a viable design for isolating the effects of target practice in a search task. However, we do not agree that the former design represents a “purer” test of the contribution of target repetitions to practice effects than our design, which compares a “repeat everything” condition with a “repeat everything but the targets” condition. In fact, we had two main reasons for using our design. Firstly, the usual situation in assessment practice is the “repeat everything” condition, for which the amount of practice benefits is already known, and we found it natural to analyze the effects of target repetitions to this usual condition. Secondly, as mentioned by Reviewer 2, one aim of our work is to identify ways for reducing practice effects in d2-like tests, and our design allows for directly testing which manipulations reduce practice effects. In addition, we would like to add that Le Dantec et al. (2012) observed comparable practice effects when “repeat targets or distractors” conditions were compared to a “repeat targets and distractors” condition, and when “repeat targets or distractor” conditions were compared to a “change targets and distractors” condition (we estimated practice effects in RTs from Table 1 in Le Dantec et al., 2012). These findings demonstrate that the two research designs can reveal similar estimates of practice effects. In the revised version of the manuscript, we are discussing the design issue in section “methodological issues” in the GD. (b) With regard to the second remark, which addressed the possibility that improvements are due to possible interactions of targets with distractors (or positions) in the “target-repetition condition,” we would also like to refer to Le Dantec et al. (2012) who did not find interactions of the effects of target practice, distractor practice, and configuration practice.

1. Given that the two groups of participants were tested together in large groups and that they were (all?) told about the changes for some participants (p. 16), I wonder whether demand characteristics could have played a role. The authors could discuss this in the General Discussion as I believe it applied to all three experiments.

Author response: It is true that participants from both conditions were tested together in Experiments 1-3. However, only participants in Experiment 1 were explicitly informed about the experimental manipulation when being instructed for the second session. In contrast, participants in Experiments 2 and 3 were not informed about the experimental manipulation. In the revised manuscript, we discuss possible effects of demand characteristics in a new section of the General Discussion, entitled “methodological issues”.

1. Tables 1-3: I was not sure what the numbers mean (probably absolute numbers of hits/KL, false alarms, and misses?), but I strongly suggest to present the data as percentages (maybe in addition to the absolute numbers) in order to allow for easier comparisons across experiments (the tasks in the three experiments differ slightly in the maximum number of targets and distractors).

Author response: In the revised version of the manuscript, the table captions better explain what the numbers in the tables mean. Moreover, in order to allow for easier comparisons across experiments, we additionally present hits as percentages, computed in relation to the total number of inspected items in session 1.

1. Figures 1-3: Error bars should be added (e.g., showing standard errors of the mean).

Author response: Done.

1. It seems that a pilot study was conducted to avoid ceiling effects in Exp. 2 (p. 23), but not for the tasks in the other two experiments. The authors may want to add an explanation for this. I assume, maybe the tasks had already been used in other studies? Also, there seems to be additional variation from session 1 to session 2 in Exp. 2 (with different spatial arrangements of the stimuli in the versions A1/B1 vs A2/B2), which was not there in Exp. 1 and 3. Why was this additional variation introduced, and could it account for some of the discrepancies in results?

Author response: We did not conduct a pilot study for Experiment 1 because the tasks used in this experiment had already been used in a previous study (Wühr, 2019), where ceiling effects were rare. Moreover, since ceiling effects were rare in Experiments 1 and 2, we did not conduct a pilot study for Experiment 3 because the test sheets used in Experiment 3 were like those used in Experiments 1 and 2, with regard to targets and number of stimuli per line. We added a corresponding statement to the revised manuscript. We changed the spatial arrangement of stimuli between sessions as an additional means for preventing ceiling effects in session 2, since we did not know the negative results of Experiment 3, with regard to configurational learning, when conducting Experiment 2. We added a corresponding statement to the revised manuscript.

1. Looking at the specific stimuli that were used as targets and distractors in the different experiments, I wonder whether the similarity between session 1 and 2 (in the change condition) was higher in Exp. 2 than in Exp. 1. That is, the target set changed from [“d d” ‘d’] to [“p p” ‘p’] in Exp. 1 (i.e., a completely new target letter), whereas the distractor set changed from [‘d "d’ ‘p’ “p p”] to [d’ "d’ ‘q’ “q q”] in Exp. 2 (i.e., quite some overlap in distractor characters). So could it be the case that the smaller difference between the two conditions in terms of practice effects was (partially) due to differences in stimulus similarity? I think the issue of similarity and generalization deserves some discussion. For instance, on p. 33, the authors state that “repeating distractors seems to play a minor role”. I am not entirely convinced by this statement as the similarity of the two sessions in Exp. 2 may have been higher than in Exp. 1, and this could at least have contributed to the differences in practice effects between Exp. 1 and 2. The authors may want to discuss this a bit.

Author response: The reviewer is basically correct in pointing out that the changes in the stimulus sets in Experiments 1 and 2 were not equivalent in terms of stimulus similarity, and therefore, might have affected the results. Lesser similarity of targets in subsequent sessions of Experiment 1 could have allowed for lesser transfer of target learning, and thus produced a relatively large performance difference between conditions in the second session of Experiment 1. In contrast, more similarity of distractors in subsequent sessions of Experiment 2 could have allowed for more transfer of distractor learning, and thus produced a relatively small performance difference between conditions in the second session of Experiment 2. Hence, comparing the size of practice benefits between these experiments might be compromised by different changes of stimulus similarities in the critical change conditions. In the revised version of the manuscript, we are discussing the issue of stimulus similarity in the discussion section of Experiment 2 and in a new section (on methodological issues) in the GD.

Minor points

p. 9, “Hence, its unclear whether” -> “Hence, it is unclear whether”

Author response: done.

p. 20, bottom, “KL” comes a bit surprising. It has been introduced early in the paper (p. 2), but not used in the more recent results section. I think the authors should use only one term (hits or KL) consistently throughout the manuscript.

Author response: In the revised manuscript, the term “hits” is used consistently throughout the manuscript.

p. 25, looking at Fig. 2, I wondered whether the pre-test difference in Hits was significant? The same is true for the difference in false alarms and misses during session 1 (Table 2).

Author response: Neither of the between-group differences in session 1 of Experiment 1 was significant: U = 1064, p = .114, for hits; U = 1022, p = .055, for false-alarms; and U = 1147, p = .306, for misses.

p. 28, line 4: Maybe I am wrong, but I think the three targets should be: “d” with two dashes above, “d” with two dashes below, and a “d” with one dash above and one below. Otherwise there seems to be overlap with the distractors.

Author response: No, you are right. We have corrected the description of the targets used in Experiment 3.

p. 28 middle paragraph: Why was one row repeated more often than the other two rows? Couldn’t the authors just have used 15 rows (5 repetitions of each row)? As it currently is, participants practiced the first row more than the second and third row. Is there any indication in the data of a stronger performance benefit (in the position repetition condition) for the first row?

Author response: We wanted to keep our test as similar as possible to the original d2 test, and therefore we provided one line of practice, which was not analyzed, and fifteen lines with test items to have five repetitions of each different spatial configuration in the analysis. As you are correctly pointing out, this construction of the test has the consequence of one spatial configuration (used in line 1) being repeated more often than the two other configurations. However, if five repetitions do not produce an advantage over the no-repetition condition, it seems unlikely to observe a difference between five and six repetitions. Nevertheless, we performed additional analyses of performance in the last three lines from session 2. In the position-repetition condition, hits slightly increased from line 14 to line 16 (6.5, 6.1, 6.7), but so did hits in the position-change condition (6.6, 6.3, 7.1). In a two-factorial ANOVA, the main effect of line was significant, F(2, 218) = 10.16, p < .001, whereas the main effect of condition (repetition, change), F(1, 109) = 1.18, p = .28, and the two-way interaction, F(2, 218) = 0.25, p = .78, were not significant. In contrast to hits, errors increased from line 14 to line 16 (1.1, 1.3, 1.4) in the position-repetition condition, whereas errors decreased (1.1, 1.0, 0,9) across the last lines in the position-change condition. In a two-factorial ANOVA, neither main effect (line: F = 0.18, p = .83; condition: F = 1.5, p = .22) nor the interaction (F = 2.13, p = .13) were significant.

The authors often refer to “numerical” differences, as opposed to “significant” differences (e.g., top of p. 26). I think this may be misleading in some places, in particular in the discussion section (such as in the discussion of Exp. 3, p. 31): I don’t think it is important to discuss a small difference in improvements when it is not significant (it would be quite unlikely to not observe a small “numerical difference” with finite samples).

Author response: We consider it useful to describe both “numerical” and “significant” differences for each experiment in order to highlight the consistency of significant and non-significant practice effects across measures. However, in the General Discussion, we are only discussing significant differences.

p. 32, third line of second paragraph: “this learning improves performance” (s missing after “improve”)

Author response: done.

p. 36: Attentional improvements and perceptual learning could also be mentioned as possible variables contributing to stimulus-independent learning.

Author response: If attentional improvements are defined as practice-dependent changes in attentional weights for targets and distractors, we would consider these improvements as stimulus-dependent learning. Similarly, if perceptual learning is defined as practice dependent changes in detecting and identifying targets, or as practice dependent changes in discriminating targets from distractors, we would consider these improvements as stimulus-dependent learning.

**References (that do not occur in the manuscript)**

Carter, E. C., et al. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*(4), 796–815. https://doi.org/10.1037/xge0000083.supp

Etherton, J. L., et al. (2018). Bayesian analysis of multimethod ego‐depletion studies favours the null hypothesis. *British Journal of Social Psychology*, *57*(2), 367–385. https://doi.org/10.1111/bjso.12236

Hagger, M. S., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Vohs, K. D., et al. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*, *32*(10), 1566–1581. https://doi.org/10.1177/0956797621989733

Wimmer, M. C., Stirk, S., & Hancock, P. J. B. (2017). Ego depletion in visual perception: Ego-depleted viewers experience less ambiguous figure reversal. *Psychonomic Bulletin & Review*, *24*(5), 1620–1626. https://doi.org/10.3758/s13423-017-1247-2

Wimmer, M. C., Dome, L., Hancock, P. J. B., & Wennekers, T. (2019). Is the letter cancellation task a suitable index of ego depletion? Empirical and conceptual issues. *Social Psychology*, *50*(5–6), 345–354. https://doi.org/10.1027/1864-9335/a000393

**Editor First Decision**: Provisional Acceptance

Feb 2, 2023

Dear Peter Wühr,

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, I would like to give you the opportunity to consider the two points mentioned by Reviewer 1 (i.e., practice effects, aim of study). I think these are helpful points, but would leave it up to you to what extent you want to incorporate them into the final version of your manuscript.

I look forward to receiving your 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.

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

Sincerely,

Jan Philipp Röer

# 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 revised manuscript has been thoroughly reviewed and revised by the authors in response to the feedback provided by reviewers, including myself. Although there may be some points of disagreement, I respect the authors’ perspective and believe that the manuscript provides valuable information on cancellation-based mental tests. I have a few final comments, but overall, I would recommend this work and found it a worthy contribution to the field.

some final comments

general. as it stands, this work is an empirical paper because it deals with experimental work but it is also a theoretical work that critically examines the many aspects of experimental potential to contribute to psychometric research.

p. 5, line 1-6, methods of dealing with practice effects
Hagemeister’s effort to detect practice effects is problematic as it requires a high level of specificity in the test to avoid false positive detection of prior test experience. I would suggest mentioning Schumann’s (2022) argument on this occasion: controlling or minimizing test-retest practice effects can be effectively done in arithmetic-based (Düker-type) tests by generating a large number of distinct items, which can be precisely adjusted in terms of difficulty (e.g. problem size). The two critical levers to control practice are: increasing the item set size and having some degree of variation in item difficulty within the test, meaning a certain degree of heterogeneity. So in summary, a test that contains a large number of distinct items within a certain range of difficulty is superior to minimize practice gains and test reliability. I wonder whether and to what degree this makes sense in a cancellation based tests as there are some differences between the two basic types of tests.

p. 6, aim of study
a comment to the response to my comment. 🙂
According to the authors, the goal of the experiments is to manipulate targets, distractors, and stimulus configurations to evaluate their impact on performance aspects. I argued that varying these aspects directly via a cognitive-experimental approach automatically implies whether is affects not only performance globally, but how psychometric aspects like test reliability are affected by experimental approach, since a psychometric test is primarily designed for the purpose of personnel selection and classification, and any cognitive-experimental study should aim to optimise psychometric criteria. I wonder if the authors would agree with this proposal. I would suggest on this occasion to consider the issue of psychometric aspects not as a side-show but as an inherent auto-implied goal of any research at the junction of cognitive-experimental and psychometric research.

appendix, inter-correlation matrix
the results are interesting, particularly the last table shows the test reliability of the primary measure (hits) from r = .70 increasing to r. = 88\*\*, I consider this enormously interesting as it shows the effect of the final experimental design varying stimulus configurations on psychometric aspects, this should be given more emphasis.

A last note on the combined performance measure in the d2 test
the d2 test delivers a combined measure that takes accuracy into account. In the Schumann (2022) study, the combined measure was even slightly more reliable, in numbers the most reliable measure of the tests. In wonder whether this would also be verified in the present context?

Schumann, F. et al. …Huestegge, L. (2022). On doing multi-act arithmetic: A multitrait-multimethod approach of performance dimensions in integrated multitasking. Frontiers in Psychology. doi:10.3389/fpsyg.2022.946626

# 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 authors have done very good job with their revision, and all my previous concerns or comments have been addressed. The manuscript, which has been in quite good shape already, has clearly improved through the revisions. I have no further suggestions.

**Author Response**
Feb 8, 2023

Dear Dr. Röer:

Thank you very much for providing quick and concise feedback on our revised manuscript for Collabra: Psychology.  We are very happy that you were now able to accept our paper for publication in Collabra: Psychology. We are confident that our paper may be of interest to readers from different areas, including experimental psychology, and psychological assessment.

In you action letter, you offered us the opportunity to make further changes to our manuscript in response to some additional comments by Reviewer 1 on the revised version of our manuscript. In fact, we have carefully thought about the comments, and made some (small) changes to our manuscript in response the reviewers’ comments. We are sending the second (final) revision of our manuscript as an attachment to this letter through the Scholastica submission site. The (new) changes to the manuscript have been highlighted by red font color. Below you will find our responses to each comment made by Reviewer 1.

Sincerely,

Peter Wühr and Bianca Wühr

**Reviewer 1**

(1) some final comments

General. As it stands, this work is an empirical paper because it deals with experimental work but it is also a theoretical work that critically examines the many aspects of experimental potential to contribute to psychometric research.

Author response: We agree.

(2) p. 5, line 1-6, methods of dealing with practice effects

Hagemeister’s effort to detect practice effects is problematic as it requires a high level of specificity in the test to avoid false positive detection of prior test experience. I would suggest mentioning Schumann’s (2022) argument on this occasion: controlling or minimizing test-retest practice effects can be effectively done in arithmetic-based (Düker-type) tests by generating a large number of distinct items, which can be precisely adjusted in terms of difficulty (e.g. problem size). The two critical levers to control practice are: increasing the item set size and having some degree of variation in item difficulty within the test, meaning a certain degree of heterogeneity. So in summary, a test that contains a large number of distinct items within a certain range of difficulty is superior to minimize practice gains and test reliability. I wonder whether and to what degree this makes sense in a cancellation based tests as there are some differences between the two basic types of tests.

Author response: We agree with Reviewer 1 and added a corresponding statement (footnote) to the Discussion of our manuscript referring to Schumann et al. (2022).

(3) p. 6, aim of study

According to the authors, the goal of the experiments is to manipulate targets, distractors, and stimulus configurations to evaluate their impact on performance aspects. I argued that varying these aspects directly via a cognitive-experimental approach automatically implies whether it affects not only performance globally, but how psychometric aspects like test reliability are affected by experimental approach, since a psychometric test is primarily designed for the purpose of personnel selection and classification, and any cognitive-experimental study should aim to optimize psychometric criteria. I wonder if the authors would agree with this proposal. I would suggest on this occasion to consider the issue of psychometric aspects not as a side-show but as an inherent auto-implied goal of any research at the junction of cognitive-experimental and psychometric research.

Author response: We do not fully agree on this point. Because we were mainly testing hypotheses, which were delineated from theories of repetition and practice effects in visual search, our study was primarily basic research, but with an eye on possible applications of our findings for improving this type of test in practice. For a study focusing on a basic-research question, the psychometric properties of the test under investigation are not directly relevant as long as the test used in our experiments is equivalent to the test used in practice, since we wish to generalize the results obtained in our experiments to the d2 and its use outside the “lab”.

(4) appendix, inter-correlation matrix

the results are interesting, particularly the last table shows the test reliability of the primary measure (hits) from r = .70 increasing to r. = 88\*\*, I consider this enormously interesting as it shows the effect of the final experimental design varying stimulus configurations on psychometric aspects, this should be given more emphasis.

Author response: We agree.

(5) A last note on the combined performance measure in the d2 test

the d2 test delivers a combined measure that takes accuracy into account. In the Schumann (2022) study, the combined measure was even slightly more reliable, in numbers the most reliable measure of the tests. I wonder whether this would also be verified in the present context?

Author response: If we understand this comment correctly, our answer would be yes. The “hit” rate is in fact a combined measure that takes accuracy into account because it is the sum of correct responses. The hit rate was always the most reliable measure in our experiments.

**Editor Final Decision:** Accept

Feb 10, 2023