Dear Prof. Claudia von Bastian,

We appreciate the positive reception from all three reviewers and your recognition of the importance of our manuscript (ID: # BR-Org-23-733). We are grateful for the constructive comments, and we acknowledge the highlighted issues that require careful attention in the revision. We addressed these concerns thoroughly and are re-submitting this revised manuscript. We believe these revisions improve the quality of this manuscript and hope that they addressed all your concerns. Please see below for our responses. For the convenience, we have highlighted changes in the revised manuscript in blue.

Chuan-Peng

On behalf of all authors

Response to the reviewers

Reviewer 1

**Reviewer Comment 1.1** — Paper Selection: First, it is unclear what were the criteria for selecting the datasets that were analysed in this study. I believe that by now probably around 100 papers have been published that investigated SPE using different forms of the matching task, while authors used the data from only 9 papers. I think that there should be a better explanation of why so few datasets were used.

Reply: Thank you for your thoughtful consideration of our paper.

To provide a detailed clarification of our selection process, we have included a flowchart, following the PRISMA diagram, illustrating the procedure used to determine the inclusion of papers in the current study (in the supplementary material, section 1.2, p.4). We hope this additional clarification addresses your concern.

In the Discussion section (p.24), we also acknowledge the limitation of analyzing a small set of papers from a larger pool. While this focused analysis enabled a deeper understanding individual-level reliability of the SPE using the SMT, we recognize that expanding the scope to include more papers could potentially bolster the generalizability of our findings.

**Figure S2.** Paper Selection Procedure (adapted from PRISMA Flow Diagram (Page et al., 2021)).

The selection of the eligible papers was based on specific criteria:

1. The paper must primarily utilize the SMT as their method.
2. The experimental design should not incorporate any stimuli that could potentially trigger a familiarity effect (e.g., using self-face, self-name).
3. The trial-level data is either openly available or shared with us upon request, enabling us to estimate at least one reliability index.

图示

描述已自动生成

**Reviewer Comment 1.1.1** — It is also a bit surprising that the datasets from the original study by Sui et al. 2012 were not included.

Reply: Thank you for bringing up a fair point about the original Sui et al. (2012) datasets not being included. Unfortunately, retrieving those specific files is not feasible due to them being stored on a university computer from over ten years ago. While replicating and building upon previous research is certainly valuable, we believe the data included in the current study provides a solid empirical basis (24 papers and 3 unpublished projects, *N* = 2250) even without those pioneering datasets.  
  
**Reviewer Comment 1.1.2** — Second, and more importantly, the selected datasets sometimes come from procedures that quite strongly diverge from the original matching task. For example, Wozniak et al 2018 used a sequential matching task in which the authors used faces and labels but presented with a 1.5 second delay between each other. And observed a bit different pattern of results (RTs effect was driven by the association of the first stimulus in the sequence, regardless of whether it was a face or a label). I think that such deviations of procedure and their influence on the results should be discussed.

Reply: Thank you for your thoughtful consideration of our study and for highlighting the potential divergence in procedures among selected datasets.

In our paper, we established inclusion criteria based on the stimuli's neutrality and the new formation of associations. Specifically, we focused on stimuli that did not involve familiarity (such as oneself or friends’ name or face) to ensure consistency across the datasets. While we acknowledge the procedural differences, such as the sequential matching task employed by Wozniak et al. 2018, we prioritized adherence to our inclusion criteria. As such, deviations from the original matching task, such as the timing of stimulus presentation or the nature of stimuli used, were considered secondary to the overarching criteria of stimulus neutrality. In the revised manuscript, we explicitly explained why we included this study. The newly added explanation is pasted below (see also, p.6):

*“For our analysis, we focused exclusively on datasets that adhered to the design of SMT without incorporating any stimuli that could potentially trigger a familiarity effect (e.g., oneself or friends’ name or face). Procedural differences from the original matching task (e.g., the timing of stimulus presentation; the nature of stimuli used), were considered secondary to the overarching criteria of stimulus neutrality.”*

To further validate our conclusions, we have conducted a supplementary analysis where we excluded the studies with procedural differences (Wozniak et al., 2018) and performed split-half reliability check. This analysis confirms the robustness of our results and demonstrates that the inclusion of studies with minor procedural variations does not change our conclusions. For instance, when the target is “Stranger”, the split-half reliability of RT was originally .49 [.35, .62], and it is .49 [.35, .61] if excluding the studies.

|  |  |  |  |
| --- | --- | --- | --- |
| Target | Indices | SHR (Original) | SHR (Excluding Wozniak et al., 2018) |
| Close | RT | .55 [.38, .70] | .55 [.39, .71] |
|  | ACC | .48 [.28, .65] | .49 [.30, .66] |
|  | *η* | .56 [.33, .74] | .57 [.34, .75] |
|  | d' | .34 [.15, .53] | .35 [.16, .53] |
|  | *v* | .29 [.01, .53] | .29 [.02, .53] |
|  | *z* | .10 [-.17, .36] | .10 [-.16, .36] |
|  |  |  |  |
| Stranger | RT | .49 [.35, .62] | .49 [.35, .61] |
|  | ACC | .48 [.34, .61] | .49 [.34, .62] |
|  | *η* | .54 [.32, .70] | .54 [.32, .70] |
|  | d' | .33 [.18, .48] | .33 [.18, .48] |
|  | *v* | .23 [.02, .43] | .23 [.02, .43] |
|  | *z* | .12 [-.08, .32] | .12 [-.08, .32] |

**Reviewer Comment 1.2**— Discussion: Several studies have already used SPE and tried to correlated it with measures reflecting individual differences. I think that such studies should be discussed, as well as whether the results obtained in this study can contribute to our understanding of these previous studies. For example: Hobbs, Sui, Kessler, Munafo, Button, 2021 for SPE and depression, Williams, Nicholson, Grainger, 2018 and Moseley, Liu, Gregory-Smith, Baron-Cohen, Sui, 2021 for autism.

Reply: Thank you for your thoughtful consideration.

We have incorporated discussions of these studies in both the introduction and the discussion sections of our manuscript (in p.3; p.22). At the group level, the interpretations remain largely consistent. However, when extrapolating to individual-level analyses, caution may be warranted. Nonetheless, reaction time (RT) measures are better, particularly in existing studies focusing on individual-level differences, where reliability results are generally higher.

Introduction, p.3: “*… in clinical investigation, the SMT has been incorporated to assess deviations in self-processing among specific populations, including individuals affected by autism or depression (e.g., Hobbs et al., 2023; Liu et al., 2022; Moseley et al., 2022). The findings from these studies are diverse. On one hand, research has demonstrated that behavioral data from SMT could function as a viable marker for depression screening (Liu et al., 2022). Additionally, performance in SMT has been employed to decode brain functional connectivity during resting state (Zhang et al., 2023) or understand the functions of self-associations in cognition (Scheller & Sui 2022a, 2023b; Sui et al., 2023; Yankouskaya et al., 2023). These studies suggest the potential for significant individual-level variability in SMT performance. On the other hand, Hobbs et al. (2023) assessed the role of self-referencing in relation to depression using SMT but found a limited association between individuals' performance in SMT and depression scores. Moseley et al. (2022) also found inconsistent correlations between SPE and its relationship to autistic traits, mentalizing ability and loneliness. These conflicting trends underscore the need to evaluate the reliability of SMT as a measurement of SPE.”*

Discussion, p.22: “…*at the group level, the interpretations of the results remain largely consistent, even without taking into account experimental parameters such as varying response rules. However, the relatively low reliability of all the SPE measures in the current analysis without considering these design parameters calls for attention when researchers are interested in individual-level analyses, such as in clinical settings or searching for an association with data from questionnaires (e.g., Hobbs et al., 2023; Moseley et al., 2022). Nonetheless, the reliability results of reaction time (RT) measures remain generally higher, particularly in existing studies focusing on individual-level differences (e.g., Liu et al., 2022; Zhang et al., 2023). Future research needs to exercise greater caution and follow the standard practice to maximize reliability at the individual level in their results (Parsons et al., 2019).”*

**Reviewer Comment 1.3**— Introduction: The article starts with introducing SPE in reference to the “cocktail party effect”. However, this is a very different type of self-bias than the one typically observed in the matching task. This is especially important, because some recent studies found that different types of self-biases appear to be quite independent of each other, see e.g: Nijhof et al 2020 “No evidence for a common self-bias across cognitive domains”.

Reply: Thank you for your valuable feedback regarding the introduction of our article and the distinction between self-bias observed in our study and other types of self-bias.

We now removed the reference to the "cocktail party effect" from the introduction. Additionally, we addressed the distinction between our self-bias measurement and other cognitive domains in the limitation section (in p.24).

*“This implies that further investigation is necessary to assess the robustness and reliability of other variations of the SMT, as well as other tasks used to measure SPE. This is particularly crucial given findings suggesting that different cognitive measures of self-biases may exhibit considerable independence from one another (Nijhof et al., 2020). ”*  
  
**Reviewer Comment 1.4** —Other Issues: The paper uses the term Self Perceptual Matching Task (SPMT). It is not a commonly used term to describe this task, and several authors argued that the matching task introduced by Sui et al 2012 is not a perceptual task, so it shouldn’t be described as a perceptual matching task. If the authors want to introduce a new term to the literature then probably it will be better to choose a less contentious term, perhaps just the Self Matching Task?

Reply: Thank you for your insightful comment regarding the terminology used in our paper.

In response to your suggestion, we have modified the term "Self Perceptual Matching Task (SPMT)" to "Self Matching Task (SMT)." We agree that the term "perceptual" may not accurately capture the nature of the task introduced by Sui et al. 2012, and we appreciate your clarification on this matter.

**Reviewer Comment 1.4.1** — Familiarity: the authors excluded datasets from experiments that involved presenting participant’s names. However, the label “You” is also highly familiar (it’s perhaps one of the most commonly used words in most languages). I think that the authors should explain why they think that it should be less problematic than participants’ names.

Reply: Thank you for your insightful comment regarding familiarity and its potential impact on our study.

In our study, we followed the approach established by Sui et al., 2012, which minimized the potential effects of familiarity by asking participants to acquire new self-relevance through learning. Also, Sui et al. (2012) conducted a series of control experiments that showed the effect of familiarity with labels in this paradigm was negligible.

It's worth noting that while the label "You" may indeed be considered familiar, it differs from participants' names in that it represents a generic identifier rather than a personalized stimulus. Participants' names inherently carry personal associations and semantic meanings unique to each individual, potentially introducing confounding variables that could impact the experimental results.

**Reviewer Comment 1.4.2** — Mismatching trials can be calculated either in reference to the neutral stimulus (e.g. a geometrical shape) or a label. For example, self-mismatching trials can be either trials involving a self-associated shape together with mismatching labels, or a self-referring label together with mismatching shapes. Please clarify which method of calculating the mismatching trials was used here.

Reply: Thank you for your insightful question. In our study, we mainly focused on matching trials, consistent with the approach in the original paper by Sui et al. (2012) and subsequent studies that highlighted “*the self-prioritization has been most robustly characterized by differences between the self and others in matching conditions*”. Therefore, our calculate only focused on matching conditions, except for the *d'*, the calculation of which involved both matching and non-matching conditions.

**Reviewer Comment 1.4.3** — My experience with the SPE is that while at the group level the effect reliably emerges, I also often have participants that do not show it at all, or even have RTs that are faster for a control category than the self. This makes me wonder if split-half reliability might not be related to the magnitude of the SPE. Perhaps the authors could check whether they correlate and potentially add this information in the supplementary materials.

Reply: Thank you for highlighting the potential correlation between split-half reliability and the magnitude of SPE. Actually, we’ve conducted this analysis and incorporated the results in the supplementary materials (in the section Exploratory Analysis, p.11).

*“We also explored the correlation between split-half reliability and effect size (Hedges’ g) and found mixed results. For some indices of SPE, the correlation between reliability and effect size is significant (e.g., RT, ACC, d’, efficiency with stranger as baseline), but for others (e.g., indices with close others as baseline), the correlation was not significant (see Fig. S9). This pattern was consistent with the reliability paradox (Hedge et al., 2018; Logie et al., 1996), suggesting that robust experimental effects are not always associated with robust individual difference correlations.*

A group of graphs with different colored lines

Description automatically generated

Fig. S9 Regression Analysis Between Permutated SHR and Effect Size (Hedges’ *g*) Using Different SPE Measures. Note: The vertical axis represents permutated split-half reliability, and the horizontal axis represents the effect size (Hedges’ *g*). Each facet represents one SPE measure.

Reviewer 2

**Reviewer Comment 2.1**— On pages 11 and 12, it is noted that reliabilities are weighted on the basis of the number of trials involved in each reliability value. This is bad practice and will artificially inflate the resulting reliability estimate, because reliabilities from larger numbers of trials per participant are naturally larger, and weighting on this basis will thus put more weight on higher reliabilities. This, in turn, will lead to overly confident conclusions that are not warranted. The authors should instead consider weighting reliability by the number of participants involved in the reliability value. The more participants, the more accurate the resulting reliability value is. Weighting should ideally occur on the basis of such indicators of accuracy.

Reply: Thank you for your valuable input.

We have re-evaluated our approach and have opted to use the sample size of participants for weighting instead. This modification indeed resulted in smaller reliability estimations, but the overall conclusion of this study did not change. We appreciate this suggestion and have updated the relevant sections in the manuscript accordingly.

New results:

图表, 条形图

描述已自动生成

Previous results:

A graph of a number of different types of text

Description automatically generated with medium confidence

**Reviewer Comment 2.2**— On page 11, the authors state that: “we used four approaches for splitting the trial-level data: first-second, odd-even, permutated, and Monte Carlo (Kahveci et al., 2022; Pronk et al., 2022). The first-second approach split trials into the first half and the second half. The odd-even approach split the trials into sequences based on their odd or even numbers. The permutation approach shuffled the trial order and randomly assigned trials to two halves. The Monte Carlo approach was similar to the permutation approach but iterated the process multiple times (usually thousands of times) to calculate the average and 95% confidence intervals of the split-half reliability.” These are not the names used by the authors of these texts for the metrics as they were described.

Reply: We sincerely appreciate the thorough examination of our manuscript and value your insightful observations.

Following the submission of both our manuscript and preprint version to PsyArXiv, we received feedback from one reader who also pointed out the original terminology for “permuted” and “Monte Carlo” split-half reliability was incorrect. Our further investigation revealed that the Monte Carlo split-half method may lead to inflated reliability because it uses the resampling method with replacement (Kahveci et al., 2022). Thus, we excluded this method from our analyses. We have updated the preprint to rectify the terminology and implemented corresponding changes throughout the manuscript and supplementary materials, including deviations from the pre-registration plan, to enhance clarity and accuracy.

p.13 in Analysis:” … To ensure methodological rigorousness, we used *three approaches* for splitting the trial-level data: first-second, odd-even and permutated… *The permutated approach shuffled the trial order and randomly assigned trials to two halves, iterating the process multiple times (usually thousands of times) to calculate the average and 95% confidence intervals of the split-half reliability.*”

p.15 in Deviation from Preregistration: “Finally, we *had incorrectly labelled the permutation method as Monte-Carlo in the first version of preprint. Thus, we corrected the misuse of the phrase in the updated version. Additionally, upon a thorough examination of the Monte-Carlo approach, we identified that its utilization could inflate reliability due to its psychometric properties (Kahveci et al., 2022). Consequently, we did not include this method in our analysis.*”

**Reviewer Comment 2.3**— While I believe the information provided by this manuscript is already a good contribution to the field, I do believe that there is more potential to this study that has gone untapped so far. Specifically, it would be helpful for researchers to know at which trial counts their SPMT is sufficiently reliable. I recommend using the Spearman-Brown prediction formula to artificially shrink or enlarge the trial counts for the studies examined and see at which count the reliability becomes sufficient.

Reply: Thank you for your constructive feedback and suggestions for further exploration of our study.

We have taken note of your recommendation regarding the use of the Spearman-Brown prediction formula to determine the trial counts at which the SMT achieves sufficient reliability. We have included the results of this analysis in the supplementary material, exploratory analysis section of our manuscript and briefly mentioned the results in our discussion.

p. 24 in Discussion: *“We used the Spearman-Brown prediction formula (Pronk et al., 2023) to predict the trial numbers required for different levels of reliability. The results indicated that the number of trials required for archive sufficient reliability (e.g., 0.8) varied across different SPE indices. For SPE measured by RT, approximately 180 trials are required to achieve a reliability of 0.8 (see Fig S11 for more caveats).*

图片包含 图示

描述已自动生成

***Fig. S11 Expected Trial Numbers Using Different SPE Measures.*** *Note: The vertical axis represents the expected trial numbers calculated based on the spearman-brown function, and the horizontal axis represents the expected split-half reliability. Each facet represents one SPE measure. For SPE measured by z, due to the confidence interval of the split-half reliability being below 0, it is not possible to use Spearman-Brown formula. Thus, only the weighted average split-half reliability of z was used.”*

**Reviewer Comment 2.4**— Furthermore, it would be useful if there was an explanation for which factors may have caused the differing effect sizes, e.g. for RT in Figure 3. If this information is provided, it will aid other authors in their future study design. Of course, if the authors looked into this and were unable to find any explanatory variables, this can be omitted.

Reply: Thank you for your insightful comment.

We conducted the analysis of effect sizes primarily to examine the stability of SPE at the group level. While we acknowledge the importance of exploring potential explanatory variables, such as moderators for the observed differences in effect sizes, it is essential to note that delving into moderator analysis would constitute a separate study. Such investigations would resemble a subgroup analysis in a meta-analysis and would be beyond the scope of the current study.

Therefore, we did not include the exploration of moderators for differing effect sizes in our paper. However, the difference contrast is indeed under consideration in our ongoing meta-analysis (see https://osf.io/euqmf, an updated version of the preregistration with the preliminary results will be available soon). We appreciate your valuable feedback.

**Reviewer Comment 2.5**— Minor Comments

**Reviewer Comment 2.5.1**— It is unclear to me why the authors justify the use of Hedges’ g and why they include the formula for ICC2 in the manuscript. These are both well known within the community.

Reply: Thank you for your feedback. We acknowledge the familiarity of Hedges' *g* and the ICC2 formula within the research community. However, we included these details for transparency and to ensure clarity for readers unfamiliar with these measures.

**Reviewer Comment 2.5.2**— It would help if authors noted whether the used ICC2 is one of "absolute agreement" or "consistency".

Reply: We utilized the Two-way random effect model based on absolute agreement (ICC2) within the ICC family. We have added this information for clarity in the manuscript (in p. 3, p.14).

p.3: “*The individual level consistency was examined using permutation-based Split-Half Reliability (r) and Intraclass Correlation Coefficient (ICC2, Two-way random effect model, absolute agreement) for assessing the consistency of task performance over time.*”

p.14: “*We focused on using the Two-way random effect model based on absolute agreement (ICC2) within the ICC family (Chen et al., 2018; Koo & Li, 2016; Xu et al., 2023).”*

Reviewer 3

**Reviewer Comment 3.1**— While I understand this is a behavioral methods papers, a bit more context here and there would be useful.  For example, on page 3, line 21: “Like other cognitive tasks… , the SPTM has been incorporated to assess deviations in self-processing”.   What did these studies show? Where they able to find evidence for individual differences? Are there examples of studies that failed to find individual differences in SPE?   I think making the context more explicit here can benefit the motivation for your study.

Reply: Thank you for your constructive feedback. We have carefully considered your suggestion and have added more context in the introduction, including information on studies that have utilized the SPMT to assess individual differences (in p.3).

*“…in clinical investigation, the SMT has been incorporated to assess deviations in self-processing among specific populations, including individuals affected by autism or depression (e.g., Hobbs et al., 2023; Liu et al., 2022; Moseley et al., 2022). The findings from these studies are diverse. On one hand, research has demonstrated that behavioral data from SMT could function as a viable marker for depression screening (Liu et al., 2022). Additionally, performance in SMT has been employed to decode brain functional connectivity during resting state (Zhang et al., 2023) or understand the functions of self-associations in cognition (Scheller & Sui 2022a, 2023b; Sui et al., 2023; Yankouskaya et al., 2023). These studies suggest the potential for significant individual-level variability in SMT performance. On the other hand, Hobbs et al. (2023) assessed the role of self-referencing in relation to depression using SMT but found a limited association between individuals' performance in SMT and depression scores. Moseley et al. (2022) also found inconsistent correlations between SPE and its relationship to autistic traits, mentalizing ability and loneliness. These conflicting trends underscore the need to evaluate the reliability of SMT as a measurement of SPE.”*

**Reviewer Comment 3.2**— In the Discussion on page 18, again a bit more context would be nice. For example, the ACC, d’ and the DDM measures proved less reliable in measuring SPE, compared to RT and efficiency. Is there evidence from other paradigms (e.g., flanker, Stroop task) that has shown similar divergence in different measures for reliability?  Do the authors have any ideas as to where these differences come from?  For example, if I understand correctly, this suggests that for the less reliable measures as ACC and drift-rate, there would less individual variability compared to RT? That makes sense too, as RT is available for every trial, but for drift rate for example, which is based on the average slope of the evidence accumulation process across many trials, there would be less data samples, hence less variability?

Reply: Thank you for your valuable feedback and suggestions for enhancing the discussion section of our manuscript.

We have taken note of your recommendation and have incorporated additional discussion regarding the reliability outcomes of measures in previous studies. We discussed the divergence in reliability compared to Reaction Time and efficiency measures and its potential implications for individual variability (in p.21). Regarding the reliability of model parameters, we also include results compared with existing research (in p.23). We hope that future studies will delve deeper into this aspect to provide further insights.

p.21: *“…SPE as measured by Reaction Time and Efficiency were higher for both split-half and test-retest reliability than other measures of SPE. These findings align with prior research (e.g., Hughes et al., 2014; Draheim et al., 2016), which also found greater within-session reliabilities for Reaction Time and accuracy composition compared to only incorporated accuracy. This is not surprising, as the difficulty of many cognitive tasks is low, making it more appropriate to focus on reaction time or a combination of reaction time and accuracy (e.g., efficiency). Similarly, the findings for the d-prime score are consistent with research on the reliability of other cognitive tasks (e.g., the matching task by Smithson et al., 2024; the recognition tasks by Franks and Hicks, 2016). It has been proposed that d-prime is heavily influenced by task difficulty, the nature of the target, and attentional factors (Vermeiren & Cleeremans, 2012). Therefore, researchers should consider these factors when using d-prime to study individual differences.”*

p.23: *“…Previous studies also found that the standard drift-diffusion model did not fit the data from the matching task (Groulx et al., 2020). Additionally, the reliability of parameters derived from other cognitive models, such as reinforcement learning models (Eckstein et al., 2022), has also been found to be unsatisfactory. These findings called for a more principled approach when modelling behavioral data to more accurately capture the fundamental cognitive processes at play (e.g., Wilson & Collins, 2019), instead of applying the standard models blindly.”*  
**Reviewer Comment 3.3**— Minor

**Reviewer Comment 3.3.1**— Transition between pg. 5/ 6. “Out of those 6 requests, 3 papers provided us with [useable] trial-level data”.  I would stop this paragraph here.

Reply: Thank you for your feedback. We've made the suggested change.

**Reviewer Comment 3.3.2**— I am not sure the details regarding the other papers are relevant at this point.   You might have done this already, but it might be useful to follow up with Bukowski et al. and let them know the data directory was empty?

Reply: Thank you for your feedback regarding the relevance of details concerning other papers.

We understand your concern and agree that specific details regarding other papers may not be directly relevant to the current discussion. As such, we refrained from mentioning them explicitly in the revisions.

Regarding our communication, we want to assure you that we have already taken this step and have initiated contact via email to address the issue. Thank you for bringing this to our attention.

Reviewer Comment 3.3.2— Page 8- 2.4 Analysis- the link in the first line here is incorrect.

Reply: Thank you for bringing this to our attention. We have corrected the link in the first line of the analysis.   
  
Reviewer Comment 3.3.3— Page 20, Conclusion, line 27: “Meanwhile, the reliability of the most robust SPE measures fell short of being satisfactory”. Can you be more specific here, and explicitly mention which measures?

Reply: Thank you for your feedback. We’ve modified the sentence to specify the measures in question: " *Meanwhile, the reliability of all the SPE measures (Reaction Time, Accuracy, Efficiency, sensitivity score, drift rate and starting point) fell short of being satisfactory.*"