BEHAVIOR RESEARCH METHODS

A Multiverse Assessment of the Reliability of the Self Matching Task as a Measurement of the Self-Prioritization Effect

Journal:	*Behavior Research Methods*
Manuscript ID	BR-Org-23-733.R1
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	28-Jun-2024
Complete List of Authors:	Liu , Zheng ; Nanjing Normal University, School of Psychology; The Chinese University of Hong Kong - Shenzhen, School of Humanities and Social Sciences Hu, Mengzhen; Nanjing Normal University, School of Psychology Zheng, Yuanrui; Nanjing Normal University, School of Psychology Sui, Jie; University of Aberdeen, Psychology Chuan-Peng, Hu; Nanjing Normal University, School of Psychology

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. The selection of the eligible papers was based on specific criteria (p.6):

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.

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 (Figure S2 in the supplementary material, section 1.2, p.4, see below). We hope this additional clarification addresses your concern.

In the Discussion section (p.24), we 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.

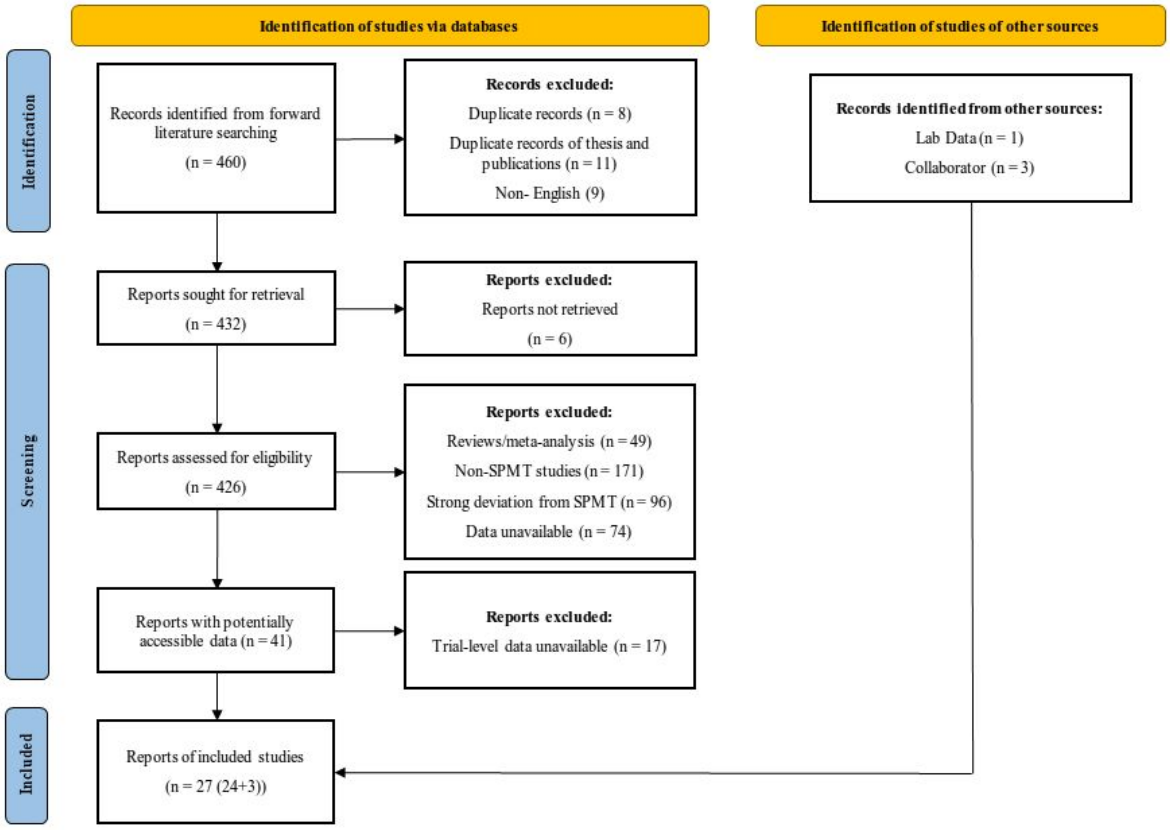


Figure S2. Paper Selection Procedure (adapted from PRISMA Flow Diagram (Page et al., 2021)).

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.

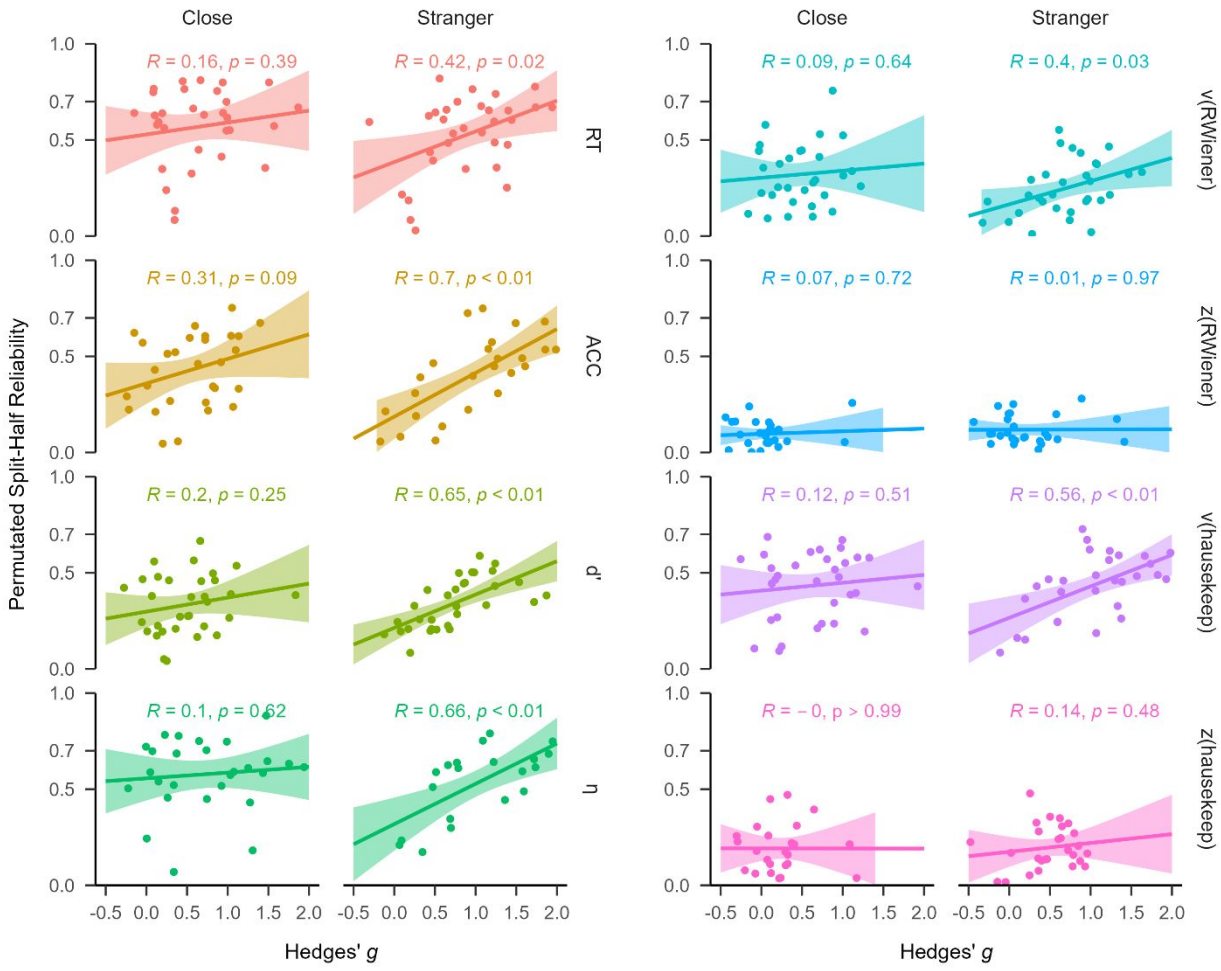


Fig. S9 Regression Analysis Between Permutated SHR and Effect Size (Hedges' g) Using Different SPE Measures. Note: The vertical axis represents permutated split-half reliability, and the horizontal axis represents the effect size (Hedges' g). Each facet represents one SPE measure.

Reviewer 2

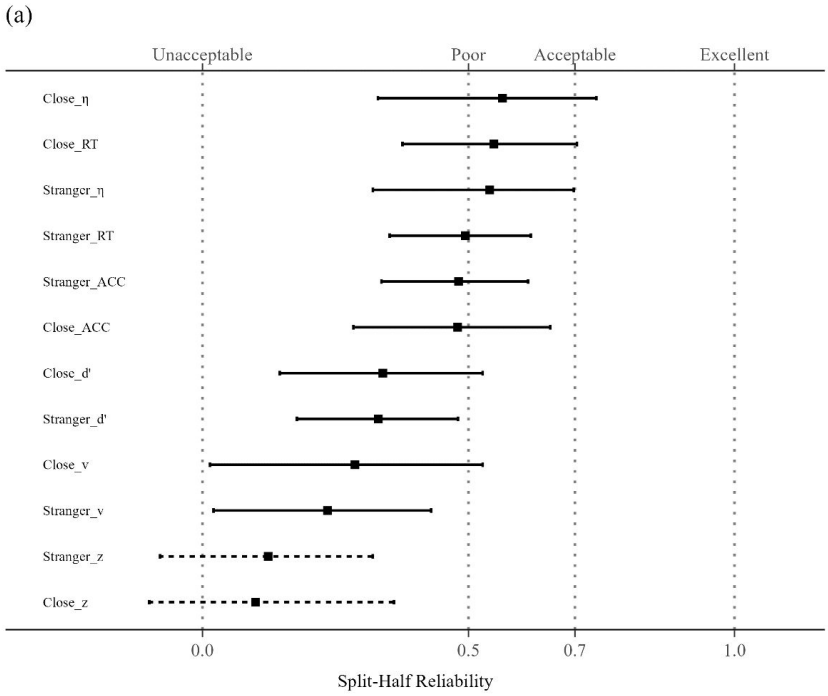
Reviewer Comment 2.1— On pages 11 and 12, it is noted that reliabilities are weighted on the basis of the number of trials involved in each reliability value. This is bad practice and will artificially inflate the resulting reliability estimate, because reliabilities from larger numbers of trials per 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.

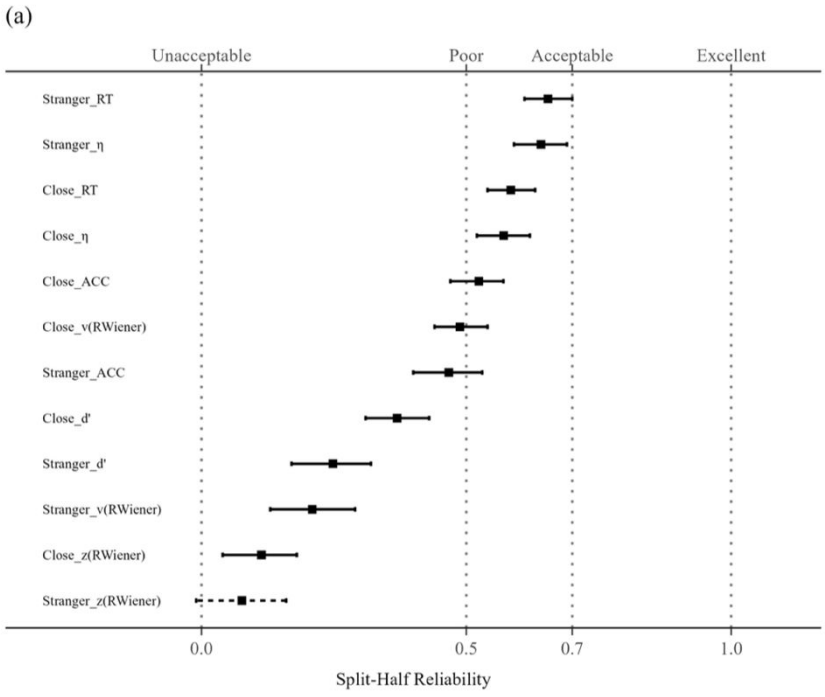
We have re-evaluated our approach and have opted to use the sample size of participants for weighting instead. This modification indeed resulted in smaller reliability estimations, but the overall

conclusion of this study did not change. We appreciate this suggestion and have updated the relevant sections in the manuscript accordingly.

New results:



Previous results:



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).*

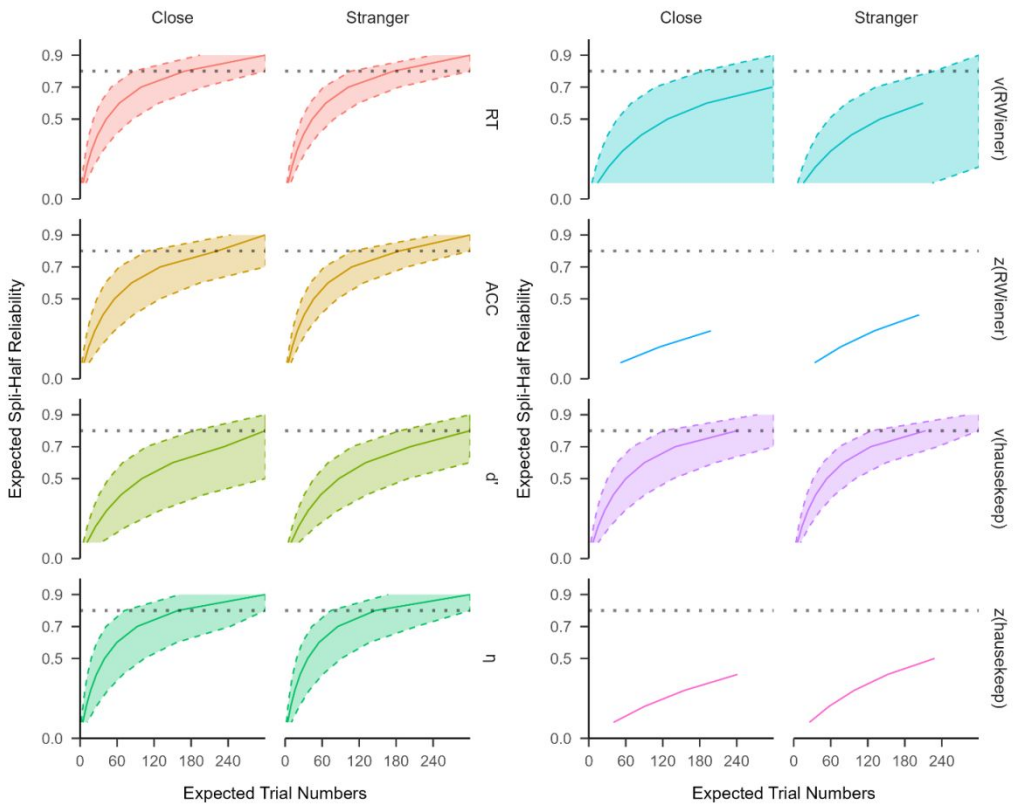


Fig. S11 Expected Trial Numbers Using Different SPE Measures. *Note: The vertical axis represents the expected trial numbers calculated based on the spearman-brown function, and the horizontal axis represents the expected split-half reliability. Each facet represents one SPE measure. For SPE measured 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.*"

A Multiverse Assessment of the Reliability of the Self Matching Task as a Measurement of the Self-Prioritization Effect

Zheng Liu[1,2†], Mengzhen Hu[1†], Yuanrui Zheng[1], Jie Sui[3], Hu Chuan-Peng[1*]

[1*]School of Psychology, Nanjing Normal University, Nanjing, China.

[2*]School of Humanities and Social Science, The Chinese University of Hong Kong-Shenzhen, Shenzhen, China.

[3*]School of Psychology, University of Aberdeen, Old Aberdeen, Scotland.

[*]Corresponding author(s). E-mail(s): hu.chuan-peng@nnu.edu.cn; hcp4715@hotmail.com;

[†]These authors contributed equally to this work.

Abstract

The Self Matching Task (SMT) is widely used to investigate the cognitive mechanisms underlying the Self-Prioritization Effect (SPE), wherein performance is enhanced for self-associated stimuli compared to other-associated ones. Although the SMT robustly elicits the SPE, there is a lack of quantifying the reliability of this paradigm. This ignorance is problematic, given the prevalence of the reliability paradox in cognitive tasks: many well-established cognitive tasks demonstrate relatively low reliability when used to evaluate individual differences, despite exhibiting replicable effects at the group level. To fill this gap, this preregistered study investigated the reliability of SPE derived from the SMT using a multiverse approach, combining all possible indicators and baselines reported in the literature. We first examined the robustness of 24 SPE measures across 42 datasets (N = 2250) using a meta-analytical approach. We then calculated the Split-Half Reliability (*r*) and Intraclass Correlation Coefficient (ICC2) for each SPE measure. Our findings revealed a robust group-level SPE across datasets. However, when evaluating individual differences, SPE indices derived from Reaction Time (RT) and Efficiency exhibited relatively higher, compared to other SPE indices, but still unsatisfied split-half reliability (approximately 0.5). For the reliability across multiple time points, as assessed by ICC2, RT and Efficiency demonstrated moderate levels of test-retest reliability (close to 0.5). These findings revealed the presence of a reliability paradox in the context of SMT-based SPE assessment. We discussed the implications of how to enhance individual-level reliability using this paradigm for future study design.

***Keywords*:** Self-Prioritization Effect (SPE), Self Matching Task (SMT), Reliability, Multiverse

1 Introduction

The Self-Prioritization Effect (SPE) reflects individuals' biased responses towards self-related information in comparison to information related to others. This phenomenon holds a central position within cognitive psychology and underscores a core facet of human cognition and self-awareness (Sui & Humphreys, 2017). SPE has been found in a broad range of cognitive tasks (e.g., Cunningham et al., 2008; Rogers et al., 1977; Sui et al., 2012). Despite SPE is often argued to be a self-specific effect, it has been challenging to be disassociated from the familiarity effect. That is, the self-related stimuli, such as own faces (Keenan et al., 2000; Kircher et al., 2000; Turk et al., 2002), own voices (Hughes & Harrison, 2013; Payne et al., 2021), or own names (Constable, Rajsic, et al., 2019) are usually more familiar to participants than those other-related stimuli. To overcome such limitation, Sui et al. (2012) introduced the Self Matching Task (SMT), where the self-relatedness (and other-relatedness) was acquired in the lab. In this task, participants first associated geometric shapes with person labels (e.g., circle = you, triangle = best friend, square = stranger) and then performed a matching task, judging whether a shape-label pair presented on the screen matched the acquired relationship. A typical pattern from this task is that shapes associated with the self exhibit a processing advantage over shapes related to others. This SPE from SMT has subsequently been replicated by many researchers (Constable, Elekes, et al., 2019; Golubickis et al., 2020; Golubickis et al., 2017; Hu et al., 2020), highlighting the robustness of the effect.

The reliability of SMT as a measurement of SPE, however, has not been examined. Here, the reliability of a cognitive task refers to its ability in producing consistent results for the same person across sessions or times (Parsons et al., 2019; Zorowitz & Niv, 2023). One common method to assess reliability is the Split-Half Reliability (r), where a test is divided into two halves, and the correlation between the data from these two halves is calculated. A high correlation suggests that the test is internally consistent and measures the same construct reliably (Pronk et al., 2022). Another widely used method is Test-retest reliability, which refers to the extent to which a measurement or assessment tool produces consistent and stable results over time when administered to the same group of individuals under identical conditions (Kline, 2015). Both methods are from classical test theory in psychometrics (Borsboom, 2005), but they are less known to experimental psychologists. In experimental research, researchers focus on the robustness of experimental effects. Robustness, in this context, pertains to the extent to which a cognitive task consistently produces the same effect at the group level across various independent participant samples. For example, the "group effect" in the Stop-Signal Task refers to differences in Reaction time between different stop-signal delays (Hedge et al., 2018). An

effect is considered robust if these differences can be consistently observed in different samples performing the Stop-Signal Task.

In recent years, driven by a growing interest in employing cognitive tasks to assess individual differences, researchers have turned their attention to evaluating the reliability of cognitive tasks (e.g., Hedge et al., 2018; Kucina et al., 2023). However, existing findings have raised concerns about the reliability of many cognitive tasks (Karvelis et al., 2023; Rouder & Haaf, 2019), with a considerable body of research highlighting moderate to low-level reliability found in the cognitive task measurements (Clark et al., 2022; Enkavi et al., 2019; Green et al., 2016). For instance, Hedge et al. (2018) reported a range of test-retest reliabilities about frequently employed experimental task metrics (such as Stroop and Stop-Signal Task), with a notable prevalence of discrepancy between the low reliability for individual differences and the robustness of the experimental effects. This discrepancy, named the "reliability paradox" (Logie et al., 1996), has gained much attention in recent years. Like other cognitive tasks, SMT was also employed by researchers as a measure of individual differences in SPE. For example, a recent study examined the individual differences in SPE and how these individual differences are correlated to brain networks (Zhang et al., 2023). Likewise, 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.

Further, the variability in quantifying SPE using SMT calls for a comprehensive examination of the reliability of different SPE measures. As simple as the SMT, there are multiple approaches to quantify the SPE, encompassing various indicators and baselines. In a typical SMT experiment, two direct outcomes are generated: Reaction Time (RT) and choices. The RT and Accuracy (ACC) of choices are the two most widely used indicators of SPE. Several other indicators can be derived from these direct outcomes: Efficiency (η) (Humphreys & Sui, 2015;

Stoeber & Eysenck, 2008), sensitivity score (d-prime, *d'*) of Signal Detection Theory (Hu et al., 2020; Sui et al., 2012), drift rate (*v*) and starting point (*z*) estimated using the Drift-Diffusion Model (DDM) (Macrae et al., 2017; Reuther & Chakravarthi, 2017). In addition to the variability of indicators, SPE can be estimated by calculating the difference between self condition and different baselines. Indeed, the selection of baselines varies across studies, such as "Close other" (e.g., Friend) (Navon & Makovski, 2021; Svensson et al., 2022), "Stranger" (Constable et al., 2021; Orellana-Corrales et al., 2020), "Celebrity" (e.g., "LuXun") (Qian et al., 2020) and "Non-person" (e.g., None) (Schäfer & Frings, 2019). As a result, three pivotal questions regarding the reliability of the SMT remain unresolved: First, given the variability of indicators (RT, ACC, *d'*, *η*, *v*, *z*) and choice of baseline conditions ("Close other", "Stranger", "Celebrity", and "Non-person"), which way of quantifying SPE is the most reliable one(s)? Second, is the SMT suitable for assessing individual differences in SPE? Finally, is there a reliability paradox in the assessment of SPE using SMT? Addressing these questions is crucial for SMT-based measurements, allowing for an accurate assessment of the SPE and its applications in various domains.

To address these three questions, the present study adopted a multiverse approach to investigate the reliability of SPE measures computed using different indicators under various baseline conditions in the SMT. This was achieved by re-analyzing 42 independent datasets (N = 2250) from 24 papers and 3 unpublished projects that employed the SMT. In order to comprehensively assess the SPE measures derived from SMT, we created a "multiverse" of possible indicators (RT, ACC, *d'*, *η*, v, z) combined with various baseline conditions ("Close other", "Stranger", "Celebrity", and "Non-person"). We first assessed the experimental effect across this multiverse using meta-analysis. 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. The findings of our study provided valuable insights into the reliability of SMT and its indicators, having the potential to facilitate the future utilization of SMT in research, clinical settings, and personal performance monitoring.

2 Methods

2.1 Ethics Information

As this study is a secondary analysis of pre-existing data sourced from publicly available datasets or archived data previously collected by the author's group, informed consent and confidentiality are not applicable.

2.2 Experimental Design

Here we provided a detailed overview of the original experimental design of SMT, as described in Experiment 1 by Sui et al. (2012). The original SMT used a 2 by 3 within-subject design. The first independent variable, labelled "Matching," consisted of two levels: "Matching" and "Non-matching", indicating whether the shape and label were congruent. The second independent variable, labelled "Identity", comprised three levels: "Self", "Friend", and "Stranger", representing the corresponding identity associated with the shape.

The original SMT consisted of two stages (refer to Fig. 1). In the first stage (instructional stage), participants were instructed to associate three geometric shapes (circle, triangle and square) with three labels (self, friend, and stranger) for approximately 60 seconds. The shape-label associations were counterbalanced between participants. In the second phase (matching task), participants completed a matching task. Each trial started with a fixation cross displayed in the center of the screen for 500 ms, followed by a shape-label pairing and fixation cross for 100 ms. The screen then went blank for 800~1200 ms, or until a response was made. Participants were required to judge whether the presented shape and label matched the learned associations from the learning phase and respond as quickly and accurately as possible by pressing one of two buttons within the allowed timeframe. Prior to the formal experimental phase, participants completed a training session consisting of 24 practice trials.

After the training, participants completed six blocks of 60 trials in the matching task, with two matching types (matching/non-matching) and three shape associations, for a total of 60 trials per association. Short breaks lasting up to 60 seconds were provided after each block.

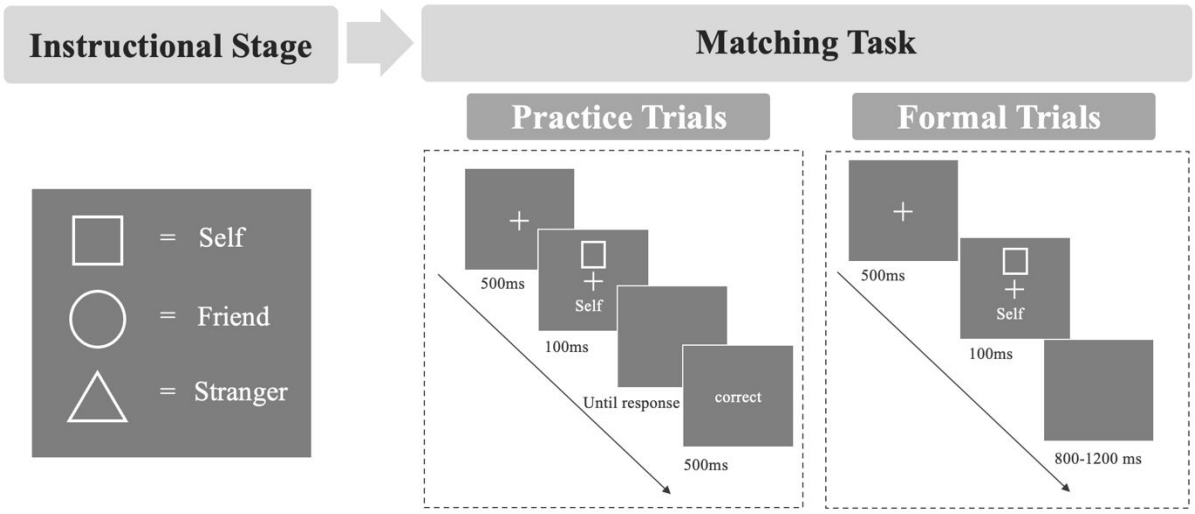


Fig. 1. Procedure of the original SMT in Experiment 1 (Sui et al., 2012). *Note*: The relation between shape-label pairs was counterbalanced between participants.

2.3 Datasets Acquisition

Initially, two datasets that employed the SMT were available to us: one from an unpublished project conducted in our laboratory (Hu et al., 2023), for which we provide more details in the supplementary materials (in section 1.1), and the other provided by our collaborators (Liu et al., 2023). Concurrently, we are conducting a meta-analysis on SPE using the SMT (Liu et al., 2021, pre-registration available at OSF: https://osf.io/euqmf). During this process, we identified an additional 24 papers with datasets potentially suitable for our present study. The detailed paper selection procedure was presented in supplementary material, Figure S2. 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.

Specifically, we identified a total of 41 papers with potentially accessible data via screening of related databases. Of these papers, 13 papers made their trial-level data publicly available. For the remaining 28 papers, we reached out to the authors and requested access to their trial-level data. Out of those 28 requests, 11 papers provided us with trial-level data. During revision, we obtained additionally two unpublished datasets (Sui 2014a, 2015). In total, our analysis

comprised raw data from 24 papers and 3 unpublished projects from our laboratory and collaborators.

It is important to highlight that the research culture discourages direct replications (Makel et al., 2012). As a result, all the datasets included in our analysis underwent some degrees of modification to the original design (e.g., change shapes, modify sequence) as well as including additional independent variables (refer to Table 1 for specification). 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; and the nature of stimuli used), were considered secondary to the overarching criteria of stimulus neutrality. For datasets from experiments that manipulated other independent variables (e.g., mood), we only utilized data from control conditions so that the data were close to the original design of SMT.

In the end, we were able to incorporate 42 independent datasets from the above-mentioned papers and projects. Nonetheless, not all studies incorporated retest sessions. If a publicly available dataset did not include a retest session with SMT, we excluded it from calculating the Intraclass Correlation Coefficient and only considered the split-half reliability. The details of the included studies and conditions in the datasets are described in Table 1.

Table 1. Dataset Information

Author & Publication Year	Study	Independent Variable				Sample Size	# of Trials per Condition	*SPE Indices*						Reliability	
		IV 1	IV 2	IV 3	IV 4			RT	ACC	*d'*	*η*	*v*	*z*	ICC	SHR
Hu et al. (2023)	1	Matching	Identity Self, Friend, Stranger	Emotion **Control**, Neutral, Happy, Sad	Session **1-6**	33	60	√	√	√	√	√	√	√	√
Constable and Knoblich (2020)	1	Matching	Identity Self, Friend, Stranger	Switch Identity Partner, Stranger	Phase **1**; 2	46	20	√	√	√	√	√	√		√
Constable et al. (2021)	2	Matching	Identity Self; Stranger	--	--	56	48	√	√	√	√	√	√		√
Qian et al. (2020)	2	Matching	Identity Self; Celebrity	Cue With, **Without**	--	25	25	√	√	√	√	√	√		√
Schäfer and Frings (2019)	1	Matching	Identity Self; Mother; Acquaintance/none	--	--	32	18	√	√	√	√	√	√		√
Golubickis and Macrae (2021)	3	Matching	Identity Self; Mother; Acquaintance	--	--	35	24	√	√	√	√	√	√		√
	1	Matching	Identity Self, Friend, Stranger	Presentation **Mixed**; Blocked	--	30	30	√	√	√	√	√	√		√
Navon and Makovski (2021)	1	Matching	Identity Self, Friend, Stranger	--	--	13	60	√	√	√	√	√	√		√
	3	Matching	Identity Self; Father; Stranger	--	--	28	60	√	√	√	√	√	√		√
	4	Matching	Identity Self, Friend, Stranger	--	--	27	60	√	√	√	√	√	√		√
Svensson et al. (2022)	1	Matching	Identity Self; Friend	--	--	20	50	√	√	√	√	√	√		√
	2	Matching	Identity Self; Friend	Frequency self > friend	--	24	100	√	√	√	√	√	√		√
	3	Matching	Identity Self; Friend	Frequency self < friend	--	25	100	√	√	√	√	√	√		√
Xu et al. (2021)	1	Matching	Identity Self, Friend, Stranger	Tasks Modified; **Unmodified**	--	105	60	√	√	√	√	√	√		√
Woźniak et al. (2018)	1	Matching	Identity Self, Friend, Stranger	Facial Gender Male; Female	--	18	56	√	√	√	√	√	√		√
	2	Matching	Identity Self, Friend, Stranger	Facial Gender Male; Female	--	18	60	√	√	√	√	√	√		√
Liu et al. (2023)	1	Matching	Identity Self; Stranger	--	--	298	16	√	√	√	√	√	√		√

Author & Publication Year	Study	Independent Variable				Sample Size	# of Trials per Condition	*SPE Indices*						Reliability	
		IV 1	IV 2	IV 3	IV 4			RT	ACC	*d'*	*η*	*v*	*z*	ICC	SHR
Sui (2014a, unpublished work)	1	Matching	Identity Self; Friend, Stranger	--	--	24	40	√	√	√	√	√	√		√
Sui (2015, unpublished work)	1	Matching	Identity Self; Friend, Stranger	--	--	20	40	√	√	√	√	√	√		√
	2	Matching	Identity Self; Friend, Stranger	--	--	21	40	√	√	√	√	√	√		√
Sui et al. (2014b)	1	Matching	Identity Self; Friend, Stranger	Frequency 1:1:1	--	24	60	√	√	√	√	√	√		√
	2	Matching	Identity Self; Friend, Stranger	Frequency 1:3:3	--	18	60	√	√	√	√	√	√		√
	3	Matching	Identity Self; Friend, Stranger	Frequency 3:1:3	--	22	60	√	√	√	√	√	√		√
	4	Matching	Identity Self; Friend, Stranger	Frequency 3:3:1	--	20	60	√	√	√	√	√	√		√
Haciahmet et al. (2023)	1	Matching	Identity Self; Stranger	--	--	40	60	√	√	√	√	√	√		√
Hobbs et al. (2023)	1	Matching	Identity Self; Friend, Stranger	--	--	144	20	√	√	√	√	√	√		√
Liang et al. (2021)	1	Matching	Identity Self; Friend, Stranger	TMS **Pre**; Post	--	109	60	√	√	√	√	√	√		√
Vicovaro et al. (2022)	1	Matching	Identity Self; Stranger	Symmetry **Symmetry**; Asymmetry	--	30	60	√	√	√	√	√	√		√
	2	Matching	Identity Self; Stranger	Symmetry **Symmetry**; Asymmetry	--	48	60	√	√	√	√	√	√		√
Woźniak & Knoblich (2022)	1	Matching	Identity Self; Friend, None	Pseudo-words Association **Pre**; Post	--	18	60	√	√	√	√	√	√		√
	2	Matching	Identity Self; Friend, None	Pseudo-words Association **Pre**; Post	--	18	60	√	√	√	√	√	√		√
	3	Matching	Identity Self; Friend, None	Pseudo-words Association **Pre**; Post	--	18	60	√	√	√	√	√	√		√

Author & Publication Year	Study	Independent Variable				Sample Size	# of Trials per Condition	*SPE Indices*						Reliability	
		IV 1	IV 2	IV 3	IV 4			RT	ACC	*d'*	*η*	*v*	*z*	ICC	SHR
Zhang et al. (2023)	1	Matching 3:1	Identity Self; Stranger	--	--	341	36	√	√	√	√	√	√		√
Sui et al. (2023)	1	Matching	Identity Self; Friend, Stranger	--	--	20	160	√	√	√	√	√	√		√
Bukowski et al. (2021)	1	Matching	Identity Self; Friend, Stranger	Training imitation; imitation-inhibition; **control-inhibition**; be-imitated	--	18	60	√	√	√	√	√	√		√
	2	Matching	Identity Self; Friend, Stranger	Training imitation; imitation-inhibition; **control-inhibition**;	--	36	60	√	√	√	√	√	√		√
Kolvoort et al. (2020)	1	Matching 1:2	Identity Self; Friend, Stranger	Delay **0ms**, 40ms, 120ms, 700ms	--	31	12	√	√	√	√	√	√		√
Martínez-Pérez et al. (2024)	1	Matching	Identity Self; Friend, Stranger	--	--	32	40	√	√	√	√	√	√		√
Feldborg et al.(2021)	1a	Matching	Identity Self; Friend	--	--	53	20	√	√	√	√	√	√		√
	1b	Matching	Identity Self; Stranger	--	--	49	20	√	√	√	√	√	√		√
Amodeo et al. (2024)	1	Matching	Identity Self; Friend, Stranger	ASD **Non-ASD**; ASD	--	30	60	√	√	√	√	√	√		√
Perrykkad & Hohwy (2022)	1	Matching	Identity Self; Friend, Stranger			286	60	√	√	√	√	√	√		√

Note. Study represents different studies from a single article; IV: independent variable. For IV3 and IV4, we only included the baseline conditions that are similar to the original design in Sui et al. (2012), which were highlighted in **BOLD font**. If other variables that could be counterbalanced are indicated by underscores, we will solely utilize these variables as stratification variables during the split-half process. Regarding the sample size, we are referring to the number of participants who meet the inclusion criteria and are therefore considered valid for the current study.

2.4 Analysis

Analysis plans for this study were preregistered on OSF (https://osf.io/zv628). All analyses in this paper were performed using the statistical software R (R Core Team, 2021). The drift rate (*v*) and starting point (*z*) of the Drift-Diffusion Model (DDM) were obtained using the "RWiener" package (Wabersich & Vandekerckhove, 2014).

The road map of the current study can be found in Fig. 2 and will be further elucidated in the subsequent sections.

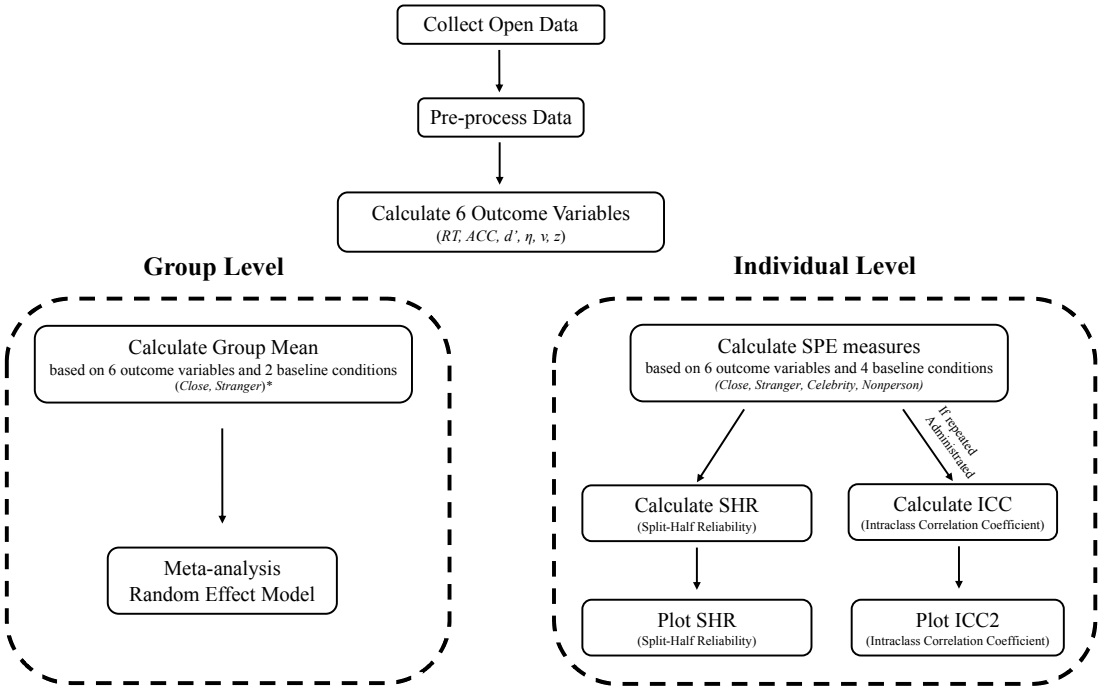


Fig. 2 Roadmap of the Current Study. *Note:* Only one paper has Celebrity and Nonpersons baseline, thus not included in the meta-analysis

2.4.1 Data Pre-processing

For all the 42 datasets (see Table 1), we applied the following exclusion criteria for excluding data:

1) Participant Exclusion Criteria
   - i Participants who had wrong trial numbers due to procedure errors are excluded from the analysis,
   - ii participants with an overall accuracy of < 0.5 are excluded from the analysis,
   - iii participants with any of the conditions with zero accuracy are excluded from the analysis.

2) Trial Level Data Exclusion Criteria
   i Trials where the keypress occurs outside the two required keys and non-responsive trials are excluded from the analysis,
   ii the practice trials are excluded,
   iii the experimental design involved independent variables more than self-referential and matching (e.g., included valence of emotion as a third independent variable).

2.4.2 Calculating the Indicators and SPE Measures

We created a "multiverse" of SPE Measures. Specifically, for each study, we first calculated six indicators for each experimental condition: Reaction Time (RT), Accuracy (ACC), Sensitivity Score (*d'*), Efficiency (η), Drift Rate (v), and Starting Point (z). Reaction Time and Accuracy were obtained directly from the datasets, while sensitivity score was calculated based on choices; Efficiency was calculated based on Reaction Time and Accuracy; Drift Rate (v) and Starting Point (z) were estimated using standard DDM with Reaction Time and choice data. As the self-prioritization effect was more often found in matching conditions (i.e., the different between self-matching condition and other-matching condition), our calculate only focused on matching conditions, except for the *d'*, the calculation of which involved both matching and non-matching conditions. The SPE Measures were then computed using different indicators under available baseline conditions in the studies (see Table. 2).

Table 2 Indicators and SPE Measures Calculation

Outcome Variables (OV)	OV Calculation	SPE Measures Calculation	Source
Reaction Times (RT)	Total Reaction Time / Total Responses	$RT_{other\text{-}matching} - RT_{self\text{-}matching}$	Sui et al. (2012)
Accuracy (ACC)	# of Correct Responses / Total Responses	$ACC_{self\text{-}matching} - ACC_{other\text{-}matching}$	Sui et al. (2012)
d-prime (*d'*)	$Z_{Hits} - Z_{False\ Alarms}$	$d'_{self} - d'_{other}$	Sui et al. (2012)
Efficiency (η)	RT / ACC	$\eta_{self\text{-}matching} - \eta_{other\text{-}matching}$	Humphreys and Sui (2015); Stoeber and Eysenck (2008)
Drift rate (v)	Parameters decomposed from RT based on standard DDM	$v_{self\text{-}matching} - v_{other\text{-}matching}$	Golubickis et al. (2017)
Starting Point (z)		$z_{self\text{-}matching} - z_{other\text{-}matching}$	Golubickis et al. (2017)

Note: OV denotes Outcome Variables; Z(.) denotes the calculation of the z-score. In this context, "hit" refers to the ACC in matching trials, while "false alarm" refers to the error rate (1 – ACC) in non-match trials; the condition "Other" varies across contrast, we calculated the SPE for each "Other" condition. More specifically, we calculated the differences for "Self vs Close", "Self vs Stranger", "Self vs Celebrity" and "Self vs Non-person".

2.4.3 Estimating the Robustness of SPE

The robustness of experimental effects (group-level effect) of SPE in SMT was calculated using a meta-analytical approach. We employed a random effects model, given the anticipated heterogeneity among participant samples (Page et al., 2021). The effect size index used for all outcome measures was Hedges' *g*, a correction of Cohen's *d* that accounts for bias in small sample sizes (Hedges & Olkin, 1985). Hedges' *g* represents the magnitude of the difference between the self and baseline condition.

When calculating Hedges' *g*, we have reversed scored the effect size for variables with negative values (Reaction Time and Efficiency). Conversely, for all indicators, a positive effect size indicates a bias towards associating stimuli with the self as compared to baseline associations. For the estimation and interpretation of effect sizes, an effect size around 0.2 was interpreted as a small effect size, around 0.5 as a medium effect size, and around 0.8 as a large effect size (Fritz et al., 2012; Hedges & Olkin, 1985).

2.4.4 Estimating the Reliability of SPE

Split-half Reliability. We assessed the split-half reliability by first splitting the trial-level data into two halves and calculating the Pearson correlation coefficients (*r*). To ensure methodological rigorousness, we used three approaches for splitting the trial-level data: first-second, odd-even and permutated (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 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.

In our analyses, we first stratified the trial-level data for each participant in the study based on experimental conditions. For example, in the case of a 2 by 3 within-subject design, we stratified the data based on the two independent variables: matching (matching, non-matching) and identity (self, stranger, friend). Subsequently, we applied the three splitting approaches (Pronk et al., 2022). When using the permutated approach, we randomly split the stratified data into two halves 5000 times, which resulted in 5000 pairs of two halves of the data. Next, we calculated 5000 Pearson correlation coefficients for these 5000 pairs. After that, we calculated the mean and 95% confidence intervals of the 5000 correlation coefficients. The first-second split and odd-even split only resulted in a single reliability coefficient. Finally, after computing the split-half reliability coefficients for each dataset, substantial variations were observed across the datasets.

To derive a more accurate estimation of the average split-half reliability for each SPE measure, we employed a synthesis approach for reliability coefficients using a minimum-variance unbiased aggregation method (Alexander, 1990; Olkin & Pratt, 1958). This approach corrects for the underestimation inherent in simply averaging correlations due to the specific distribution properties of correlation coefficients (Shieh, 2010). The method involves a correction and weighting of the reliability coefficients based on the number of participants. We calculated the weighted-average reliabilities using the "cormean" function within the "AATtools" Package (Kahveci, 2020). While there is no strict criterion for defining the level of split-half reliability for psychological and educational measures, a widely accepted guideline suggests that a value of 0.5 is considered "poor," a value of 0.70 is deemed "acceptable," and a value greater than 0.8 indicates excellent reliability (Cicchetti & Sparrow, 1981).

Test-Retest Reliability (ICC). The Intraclass Correlation Coefficient (ICC) serves as a widely recognized measure for evaluating test-retest reliability (Fisher, 1992). Differing from the Pearson correlation coefficient, which primarily quantifies the linear association between two continuous variables, the ICC extends its prowess to scenarios involving multiple measurements taken on the same subjects, while also considering both the correlation and agreement between multiple measurements, making it a more comprehensive measure of test-retest reliability (Koo & Li, 2016). Since our primary aim was to evaluate the appropriateness of the SMT in assessing individual differences and repeated administration, to achieve this objective, we assessed the test-retest reliability of the six indicators for our dataset that involved test-retest sessions using the function "ICC" in the "psych" package (Revelle, 2017). 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). ICC2 gives an estimate of the proportion of total variance in measurements that is attributed to between-subjects variability (individual differences) and within-subjects variability (variability due to repeated measurements) (Xu et al., 2023). For the calculation of ICC2 estimates, the formula is:

$$\mathrm{ICC2} = \frac{\mathrm{MS_{BS}} - \mathrm{MS_E}}{\mathrm{MS_{BS}} + (\mathrm{k}-1)\mathrm{MS_E} + \frac{\mathrm{k}}{\mathrm{n}}(\mathrm{MS_{BM}} - \mathrm{MS_E})} \qquad (1)$$

where MSBS is the mean square between subjects, MSE is the mean square error, MSBM is the mean square between measurements, *k* is the number of measurements, *n* is the number of participants.

The traditional benchmarks for interpreting ICC values are as follows: ICC less than 0.50 suggests poor reliability; ICC between 0.50 and 0.75 suggests moderate reliability; ICC between

0.75 and 0.9 suggests good reliability; ICC above 0.9 suggests excellent reliability (Cicchetti & Sparrow, 1981; Kupper & Hafner, 1989).

3 Deviation from Preregistration

We adhered to our pre-registration plan as much as possible, however, there were a few differences between the current report and the pre-registration document. First, in our initial preregistration plan, we did not anticipate analyzing the group-level effect of SPE due to the perceived robustness of the effect across a diverse range of research. However, as our study progressed, we recognized the value of providing a more comprehensive assessment. Thus, we included an estimation of pooled effect sizes across the included study to represent the group-level effect. Second, we used a different algorithm for estimating the parameters of the drift-diffusion model. In the preregistration, we planned to estimate the drift rate (v) and starting point (z) of the Drift-Diffusion Model using the "fit_ezddm" function from the "hausekeep" package (Lin et al., 2020). This function served as a wrapper for the EZ-DDM function (Wagenmakers et al., 2007). However, we observed limitations in the algorithm's ability to accurately estimate parameter z during parameters recovery (details provided in the Supplementary Materials, section 1.2). After comparing the 5 algorithms, we found that the "RWiener" package (Wabersich & Vandekerckhove, 2014) achieved a favorable balance between accuracy, confidence interval and computational efficiency, making it the most suitable choice for our analysis. Nevertheless, for transparency, we have included the results from ezDDM in the supplementary materials (see Supplementary, Fig. S2-4). Third, we did not explicitly state in the preregistration report that we would perform a weighted average of the split-half reliabilities for all datasets. However, to obtain an overall estimate of the reliability, we weighted each study based on the number of participants. Fourth, in our original preregistration, we outlined our intention to include both ICC2 and ICC2k in our data analysis. However, as our understanding of Intraclass Correlation Coefficients (ICC) improved, we realized that ICC2 is the appropriate index for our research purpose. More specifically, ICC2k was mentioned in the preregistration as an index of robustness of group-level effect, but it turned out to be another index of reliability for individual differences. We corrected this misinterpretation of ICC2k in the final report. Fifth, we conducted exploratory analysis using the data we collected to investigate the relationship between the number of trials, permutated split-half reliability, and effect size (Hedges' g) (refer to Supplementary Fig. S8-10). In addition, as suggested by one reviewer, we used the Spearman-Brown prediction formula based on our current data to predict the trial counts at which the SMT achieves sufficient reliability (Pronk et al., 2023). Sixth, the writing of the current manuscript was improved based on the pre-registration. For example, in our preregistration, we included

different baseline conditions when calculating SPE in the method section but did not mention this in our introduction and abstract. Finally, we had incorrectly labelled the permutation method as Monte Carlo in the first version of the 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.

4 Results

Of the 42 independent datasets, 34 of them contain data for "Close other", 34 of them contain data for "Stranger", 1 of them has data for "Celebrity", and 4 of them have data for "Nonperson". Since there were only a few datasets for "Celebrity" and "Nonperson", their results were presented in the supplementary materials.

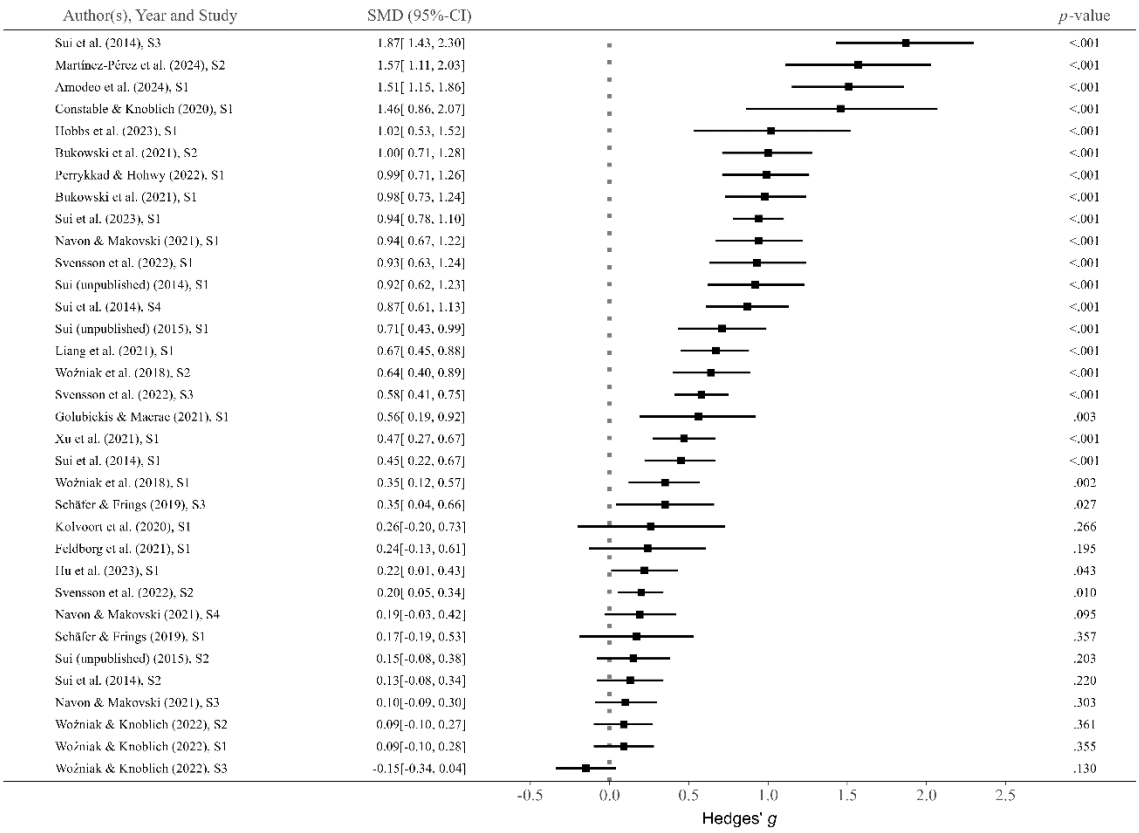
4.1 Group Level Effect of SPE

We conducted a meta-analytical assessment to examine the robustness of SPE as measured by SMT. We used a random effect model to synthesize the effect across different studies, with Hedges' g as the index of effect size. We found that all measures of SPE, except the parameter z estimated from DDM, exhibited moderate to large effect sizes (see Table. 3 for numeric results for all six SPE measures, Fig. 3 for forest plots of effect sizes for RT). Our findings indicated a robust and substantial experimental effect of SPE. The I^2 value, all being greater than 75%, indicates high heterogeneity among studies, justifying the selection of the random effect model (Borenstein et al., 2021). The results for "Celebrity" and "None" as baselines were included in the supplementary materials (see Supplementary Table. S1).

Table 3. Meta-Analytical Results of SPE Measures in SMT

Baseline	Indicators	Hedges′ *g* [95% CI]	# of Studies	Q	*p*	I^2
Close						
	RT	.61 [.45, .77]	34	363.24	<.001	92.86%
	ACC	.73 [.48, .97]	34	457.56	<.001	95.98%
	d	.48 [.34, .62]	34	297.84	<.001	91.06%
	η	.93 [.66, 1.20]	34	556.01	<.001	96.51%
	v	.49 [.36, .62]	34	315.23	<.001	90.28%
	z	-.06 [-.19, .08]	34	368.55	<.001	92.70%
Stranger						
	RT	.94 [.75, 1.13]	34	391.22	<.001	92.85%
	ACC	1.42 [1.00, 1.84]	34	667.53	<.001	97.60%
	d	.66 [.49, .84]	34	381.71	<.001	93.09%
	η	1.69 [1.23, 2.15]	34	702.01	<.001	97.89%
	v	.70 [.52, .89]	34	400.50	<.001	93.86%
	z	.05 [-.13, .24]	34	530.75	<.001	95.62%

(a) RT for [Self - Close]



(b) RT for [Self - Stranger]

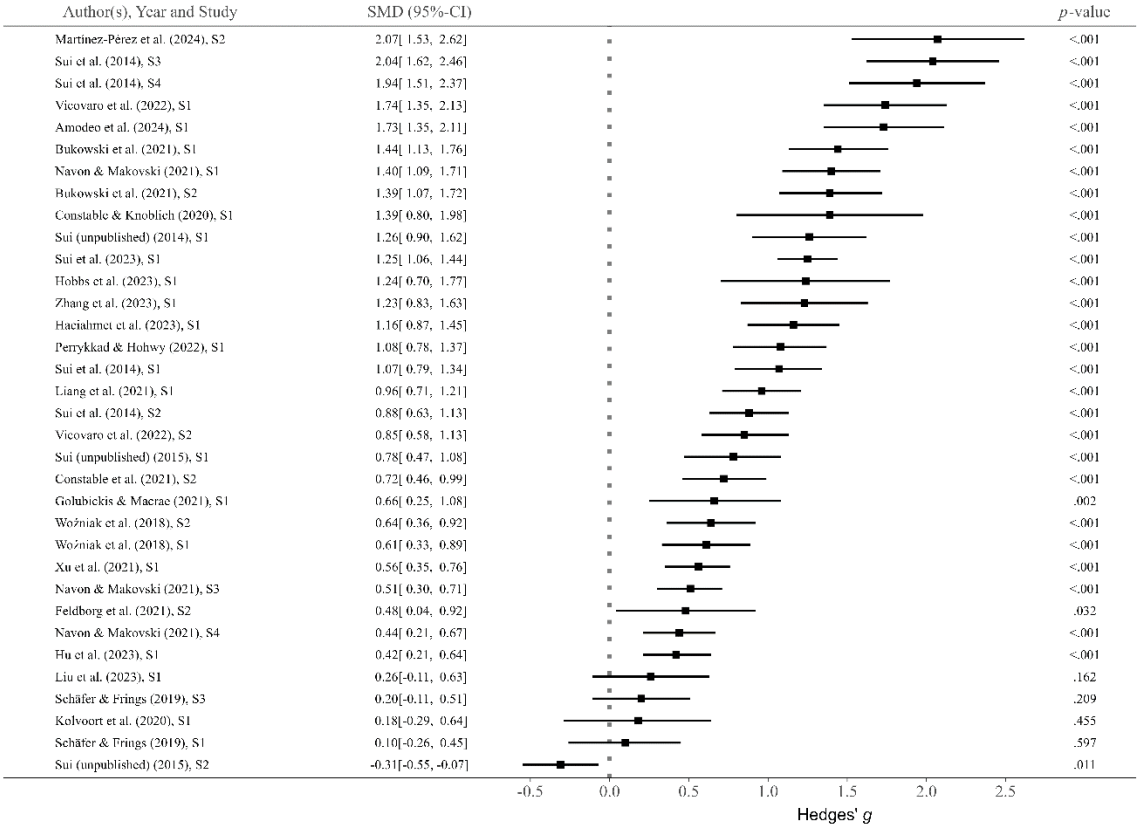


Fig. 3 Forest Plots for Group-level Self-Prioritization Effect (SPE) as Quantified by RT. (a) When "Close other" as the baseline condition for SPE, i.e., the "Self vs. Close other" contrast; (b) When "Stranger" as the baseline condition for SPE, i.e., the "Self -Stranger" contrast.

4.2 Split-half Reliability

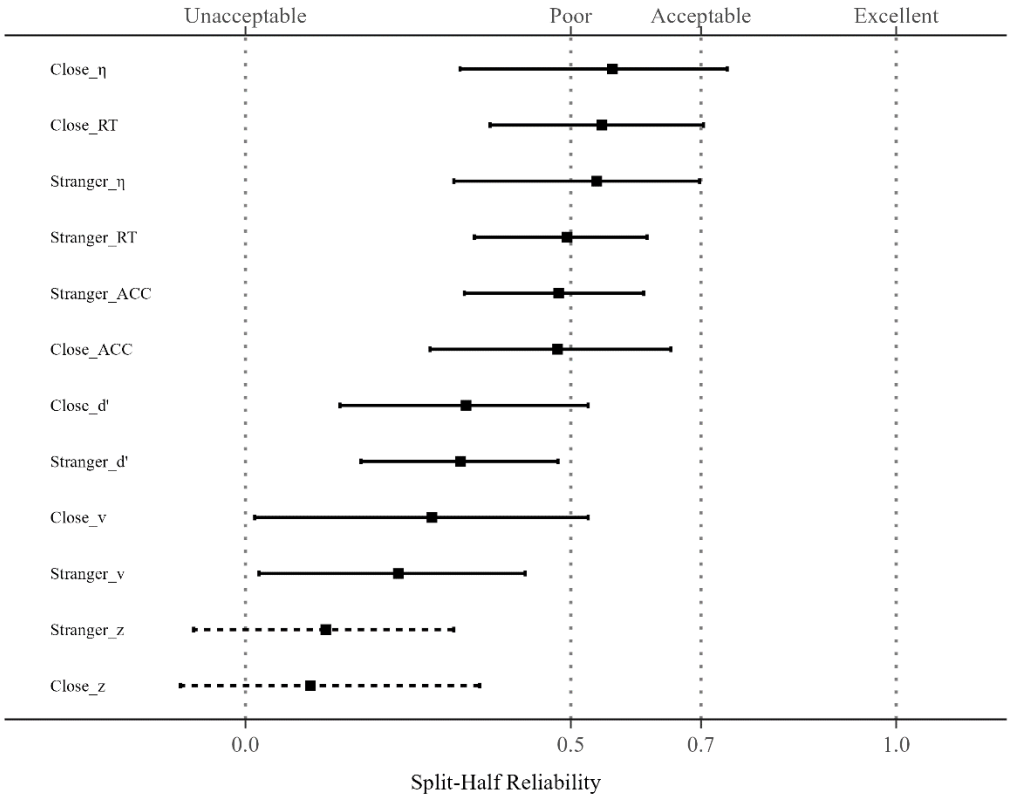
We used three different approaches to split the data when calculating split-half reliability: the first-second, odd-even and permutated methods. Also, we used the weighted average split-half reliability as the overall reliability across studies. Here we only presented the results from the permutated split-half method both for clarity and for the robustness of this approach (Pronk et al., 2022) (see Fig. 4(a)). The results of the other two split-half methods can be found in the supplementary materials (see Supplementary Fig. S5).

We found that, among all SPE measures, the four with highest split-half reliabilities were as follows: Reaction Time (RT) with "Close other" as baseline (r = .55, 95%CI [.38, .70]); Efficiency (η) with "Close other" as baseline (r = .56, 95% CI [.33, .74]); η with "Stranger" as baseline (r = .54, 95%CI [.33, .70]); RT with "Stranger" as baseline (r = .49, 95%CI [.35, .62]). These SPE measures achieved a split-half reliability of around 0.5 or higher, which is considered poor. For all other SPE measures, the reliability was around 0.5 or lower, indicating poor reliability. These included Accuracy (ACC), Sensitivity Score (d'), Drift Rate (v), and Starting Point (z) under four baselines. It's worth noting that the split-half reliability of z, the starting point parameter estimated from DDM, for all baselines was around 0, which suggested a total lack of reliability.

4.3 Test-retest Reliability

ICC could only be calculated for the dataset from our laboratory (Hu et al., 2023), which has 2 baseline conditions, the "Close other" and "Stranger", in the experimental design. The ICC2, which measures the reliability for individual differences, aligns with the findings observed in split-half reliability estimation (see Fig. 4(b)). Specifically, when using "Close other" as the baseline, the ICC2 for SPE measured by RT was .53 (95% CI [.39, .69]), and for Efficiency, it was .52 (95% CI [.38, .68]). Meanwhile, when "Stranger" was used as the baseline, the ICC2 for RT was .58 (95% CI [.45, .73]), and for Efficiency, it was .35 (95% CI [.21, .52]). All other measures of SPE exhibited reliability lower than 0.5. To test the robustness of the results, we explored one additional dataset that included a re-test session but deviated strongly from the original SMT, the result showed a similar pattern here (see Supplementary Fig. S6).

(a)


Unacceptable
Poor
Acceptable
Excellent
Close_η
Close_RT
Stranger_η
Stranger_RT
Stranger_ACC
Close_ACC
Close_d'
Stranger_d'
Close_v
Stranger_v
Stranger_z
Close_z
0.0
0.5
0.7
1.0
Split-Half Reliability

(b)

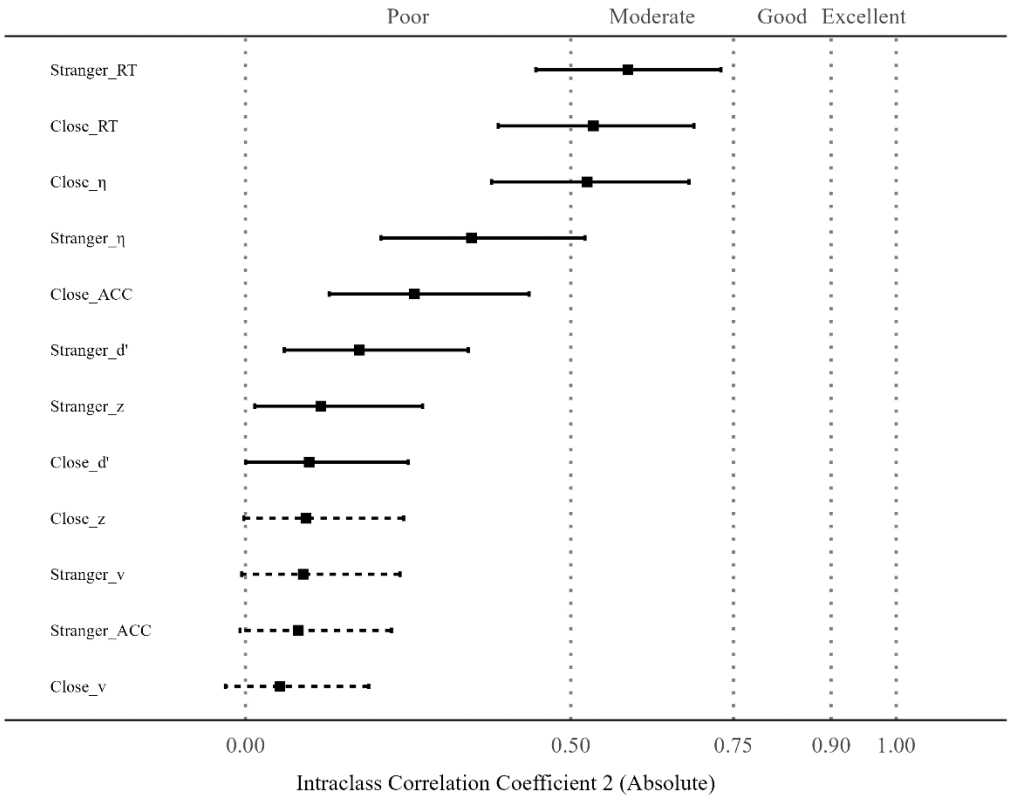

Poor
Moderate
Good
Excellent
Stranger_RT
Close_RT
Close_η
Stranger_η
Close_ACC
Stranger_d'
Stranger_z
Close_d'
Close_z
Stranger_v
Stranger_ACC
Close_v
0.00
0.50
0.75
0.90
1.00
Intraclass Correlation Coefficient 2 (Absolute)

Fig. 4. Reliability for Different SPE Measures. (a) The Weighted Average Split-Half Reliability (Permutated); (b) Intraclass Correlation Coefficient (ICC2). Note: The vertical axis represents 12 different SPE measures, combining six indicators (RT, ACC, *d'*, η, *v*, *z*) and two baseline conditions ("Close other" and "Stranger"). The weighted average split-half reliability (figure a) and ICC values and their corresponding 95% confidence intervals are illustrated using points and lines. The dashed line indicates that the confidence interval for that point estimate extends across 0, implying a lack of reliability. Since there is only one paper for "Celebrity" and four for "Nonperson", their results are presented in the supplementary materials.

5 Discussion

In this pre-registered study, we examined the reliability of various measures from the Self Matching Task (SMT) in assessing the self-prioritization effect (SPE). Our analyses revealed that except for parameters *z* from DDM, all the other measures exhibited robust SPE. However, when it came to reliability for individual differences, only two measures of SPE, Reaction Time and Efficiency, exhibited relatively higher but still unsatisfactory reliability, among all indicators that have been reported in the literature. Notably, this variability in reliability may not be taken into account for experimental parameters such as duration of stimulus presentation and response rules in the analyses - key parameters that influence cognitive task performance (Hedge et al, 2018). Our results revealed a "reliability paradox" for the SMT, when relying only on the matching task type and unfamiliar stimuli. These findings provided important methodological insights for using the SMT in assessing SPE at the individual level.

First, the Reaction Time (RT) and Efficiency (η) appeared to be the best measures among all the different ways to measure SPE (the other were ACC, *d'*, the parameter *v* and *z* from DDM). Our results revealed that the Reaction Time and Efficiency performed relatively well on both group level and individual levels. On the group level, effect sizes of SPE as measured by Reaction Time and Efficiency were moderate to large effect; on the individual level, 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'* 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'* 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'* to study individual differences. In addition, for different baseline conditions used for calculating SPE in the literature, "Stranger" and "Close other" (e.g., friends, or mother) are the most commonly utilized. Notably, "Stranger" produced a slightly higher effect size for most of the six indicators and demonstrated greater reliability when it came to Reaction Time. These results aligns with the preliminary results of our ongoing meta-analysis (Liu, et al.,2021), suggesting that the selection of a baseline could be a significant moderator of the SPE. Taken together, for researchers interested in balancing between the group-level SPE and reliability, using Reaction Time and Efficiency as the indicators might be a good choice.

Second, taking the group-level robustness and individual-level results together, our findings revealed a "reliability paradox" for SMT that were similar to the experimental settings of Sui et al (2012). We observed that the majority of the SPE measures demonstrated moderate to large effect sizes when analyzed at the group level. However, when considering individual differences, only the SPE measures derived from RT and Efficiency displayed comparatively higher values than other SPE measures but still did not meet the criteria for satisfactory split-half reliability. Likewise, when examining the reliability across multiple time points using ICC2, RT and Efficiency still ranked the highest but only showed moderate levels of test-retest reliability. Our finding also aligned with the "reliability paradox" of cognitive tasks discovered in previous studies (Enkavi et al., 2019; Hedge et al., 2018). The precise causes behind the reliability paradox observed in SPE measurements using the SMT warrant thorough investigation. However, one of the plausible explanations is that the SMT, like other cognitive tasks, tends to exhibit minimal variability among participants while maximizing the detection of SPE at the group level (Liljequist et al., 2019). Alternatively, the current finding indicates that when assessing reliability for individual differences, it is essential to consider critical experimental parameters such as stimulus presentation, response rules, stimulus onset asynchrony/inter-trial interval, and the number of practice trials. These parameters enable a fine-grained design for individual differences in SPE using the SMT. The current study sheds light on the specific types of inquiries on how to proficiently use the SMT to address both group and individual-level differences in SPE. More specifically, 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).

Many studies have previously used the SMT to assess robust group-level SPE, more recent studies showed a burgeoning interest in using the SMT to quantify individual variability in SPE. Several approaches have recently been proposed to enhance the reliability of cognitive tasks, which may prove valuable for the SMT. These include using gamification (Friehs et al., 2020, Kucina et al., 2023), latent model (Eisenberg et al., 2019; Enkavi et al., 2019) or generative models (Haines et al., 2020) to analyze the data. Some of these suggestions have already been validated by empirical data. For example, Kucina et al. (2023) re-designed the cognitive conflict task by incorporating more trials and gamification indeed improving the reliability compared to the traditional Stroop task alone. Our exploratory analyses of the relationship between trial numbers and reliability also suggest that increasing trial numbers may improve reliability (please refer to Supplementary section 2.4).

Finally, a surprising result is the notably low split-half and test-retest reliability observed in the parameters (v and z) derived from the drift-diffusion model. In our analyses, we applied common and easy-to-use methods to datasets, estimated parameters for each condition of each participant and then calculated the reliability. The reliability of both the drift rate (v) and the starting point (z) fell well below acceptable levels. These results contradict previous findings that drift rate (v) and starting point (z) can be used as an index of SPE. Several studies interpreted the drift rate (v) as the index of the speed and quality of information acquisition and reported higher drift rate for self-relevant stimuli (e.g., Golubickis et al., 2020; Golubickis et al., 2017). However, the reliability of drift rate (v) is relatively low in our study. As for the starting point (z), studies also reported SPE using z and interpreted this effect as a preference for matching response when the stimuli are self-relevant (e.g., Macrae et al., 2017; Reuther & Chakravarthi, 2017). Our meta-analytical results indicated that the Hedges' g for starting point (z) was around zero. The split-half reliability of z was also small, possibly because z fails to adequately reflect the SPE. These findings raised serious concerns about applying the standard drift-diffusion model to data from SMT directly. 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 behavioural data to more accurately capture the fundamental cognitive processes at play (e.g., Wilson & Collins, 2019), instead of applying the standard models blindly.

5.1 Implications of the Current Study

Our findings can offer an initial guide for researchers considering the use of SMT. Firstly, we recommend that researchers employ Reaction time and Efficiency as the indicators of SPE since they strike a balance between achieving a substantial effect size at the group level and ensuring reliability at the individual level. Second, if researchers are interested in a relatively bigger group-level effect size, using the "Self vs Stranger" contrast may prove beneficial. Third, if feasible, increase the number of trials, as it may enhance the overall reliability of the measurements. 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). Lastly, we caution against the careless application of the standard drift-diffusion model and instead advocate for a principled modelling approach.

5.2 Limitations

Several limitations warrant acknowledgment. Firstly, although we made efforts to enhance sample diversity by including open data when available, it is important to note that the majority of our samples still consisted of individuals from what is commonly referred to as "(W)EIRD" populations (Rad et al., 2018; Yue et al., 2023), most of the participants were recruited from universities and are healthy adults. As a result, our findings may not be fully representative of the broader population, and it is necessary to include a more diverse sample to ensure greater generalizability of the paradigm. Secondly, the results presented in this study evaluated the robustness and reliability of SPE using the SMT by Sui et al. (2012). While this focused analysis in a small set of papers from a large pool using the SMT enabled a deeper understanding individual-level reliability of the SPE, we recognize that expanding the scope and criteria to include more papers could potentially bolster the generalizability of our findings. 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). Thirdly, when assessing the intraclass correlation coefficients (ICC2), only one dataset had available data from multiple tests, which could potentially limit the representativeness of the results. This issue is mitigated by the fact that additional analysis of one dataset (see supplementary section 2.3) with different designs showed similar results as we reported in the main text.

6 Conclusion

This study provided the first empirical assessment of the reliability of the self matching task (SMT) for measuring individual differences in SPE. We found a robust self-prioritization effect for all measures of SPE, except the starting point parameter *z* estimated from DDM. Meanwhile, the reliability of all the SPE measures (Reaction Time, Accuracy, Efficiency, sensitivity score, drift rate and starting point) fell short of being satisfactory. The results of the current study may serve as a benchmark for the improvement of individual-level reliability using this paradigm.

Acknowledgments

The data collection from Hu et al. (2023) was supported by the National Science Foundation China (Grant No. 31371017) to JS. The author wishes to express gratitude to Dr. Sercan Kahveci for his valuable feedback on the first version of the preprint.

Author Contributions

HCP: Conceptualization, Methodology, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Project administration, Supervision. **LZ:** Methodology, Data Curation, Software, Formal analysis, Visualization, Investigation, Writing - Original Draft. **HMZ:** Methodology, Data Curation, Software, Formal analysis, Visualization, Investigation, Writing - Original Draft. **ZYR**: Software. **SJ**: Funding acquisition, Data Curation, Writing - Review & Editing.

Data and Material Availability

The pre-registration plan is available at OSF (https://osf.io/zv628). The de-identified raw data from our lab is available at Science Data Bank (https://doi.org/10.57760/ sciencedb.08117). The simulated data is accessible on GitHub (https://github.com/ Chuan-Peng-Lab/ReliabilitySPE).

Code Availability

The code used to simulate and analyze the data is made accessible on GitHub (https: //github.com/Chuan-Peng-Lab/ReliabilitySPE).

Competing Interests

The authors declare no competing interests.

References

References marked with an asterisk (*) indicate papers that are included in the analysis.

Alexander, R. A. (1990). A note on averaging correlations. *Bulletin of the Psychonomic Society*, *28*(4), 335-336.

*Amodeo, L., Goris, J., Nijhof, A. D., & Wiersema, J. R. (2024). Electrophysiological correlates of self-related processing in adults with autism. *Cognitive, Affective, & Behavioral Neuroscience*, 1–17. https://doi.org/10.3758/s13415-024-01157-0

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis.* John Wiley & Sons.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge University Press.

*Bukowski, H., Todorova, B., Boch, M., Silani, G., & Lamm, C. (2021). Socio-cognitive training impacts emotional and perceptual self-salience but not self-other distinction. *Acta Psychologica*, *216*, 103297. https://doi.org/10.1016/j.actpsy. 2021.103297

Chen, G., Taylor, P. A., Haller, S. P., Kircanski, K., Stoddard, J., Pine, D. S., Leibenluft, E., Brotman, M. A., & Cox, R. W. (2018). Intraclass correlation: Improved modeling approaches and applications for neuroimaging. *Human Brain Mapping*, *39*(3), 1187–1206. https://doi.org/10.1002/hbm.23909

Cheng, M., & Tseng, C.-h. (2019). Saliency at first sight: Instant identity referential advantage toward a newly met partner. *Cognitive Research: Principles and Implications*, *4*(1), 1–18. https://doi.org/10.1186/s41235-019-0186-z

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25* (5), 975–979. https://doi.org/10.1121/1.1907229

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American*

Journal of Mental Deficiency, *86*(2), 127–137. https://psycnet.apa.org/record/198200095-001

Clark, K., Birch-Hurst, K., Pennington, C. R., Petrie, A. C., Lee, J. T., & Hedge, C. (2022). Test-retest reliability for common tasks in vision science. *Journal of Vision*, *22*(8), 18–18. https://doi.org/10.1167/jov.22.8.18

Constable, M. D., Elekes, F., Sebanz, N., & Knoblich, G. (2019). Relevant for us? we-prioritization in cognitive processing. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(12). https://doi.org/10.1037/ xhp0000691

*Constable, M. D., & Knoblich, G. (2020). Sticking together? re-binding previous other associated stimuli interferes with self-verification but not partner-verification. *Acta Psychologica*, *210*, 103167. https://doi.org/10.1016/j.actpsy.2020.103167

Constable, M. D., Rajsic, J., Welsh, T. N., & Pratt, J. (2019). It is not in the details: Self-related shapes are rapidly classified but their features are not better remembered. *Memory & Cognition*, *47*, 1145–1157. https://doi.org/10.3758/ s13421-019-00924-6

*Constable, M. D., Becker, M. L., Oh, Y.-I., & Knoblich, G. (2021). Affective compatibility with the self modulates the self-prioritisation effect. *Cognition and Emotion*, *35*(2), 291–304. https://doi.org/10.1080/02699931.2020.1839383

Cunningham, S. J., Turk, D. J., Macdonald, L. M., & Macrae, C. N. (2008). Yours or mine? ownership and memory. *Consciousness and Cognition*, *17* (1), 312–318. https://doi.org/10.1016/j.concog.2007.04.003

Draheim, C., Hicks, K. L., & Engle, R. W. (2016). Combining Reaction Time and Accuracy: The Relationship Between Working Memory Capacity and Task Switching as a Case Example. *Perspectives on Psychological Science*, *11*(1), 133–155. https://doi.org/10.1177/1745691615596990

Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. (2022). The interpretation of computational model parameters depends on the context. *Elife*, *11*, e75474. https://doi.org/10.7554/eLife.75474

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven

ontology discovery. *Nature Communications*, *10* (1), 2319. https://doi.org/10.1038/s41467-019-10301-1

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

Enock, F. E., Sui, J., Hewstone, M., & Humphreys, G. W. (2018). Self and team prioritisation effects in perceptual matching: Evidence for a shared representation. *Acta Psychologica*, *182*, 107–118. https://doi.org/10.1016/j.actpsy.2017.11. 011

*Feldborg, M., Lee, N. A., Hung, K., Peng, K., & Sui, J. (2021). Perceiving the self and emotions with an anxious mind: evidence from an implicit perceptual task. *International Journal of Environmental Research and Public Health, 18*(22), 12096. https://doi.org/10.3390/ijerph182212096

Fisher, R. A. (1992). *Statistical methods for research workers.* Springer New York. https://doi.org/10.1007/978-1-4612-4380-9 6

Franks, B. A., & Hicks, J. L. (2016). The reliability of criterion shifting in recognition memory is task dependent. *Memory & Cognition*, *44*, 1215–1227. https://doi.org/10.3758/s13421-016-0633-8

Friehs, M. A., Dechant, M., Vedress, S., Frings, C., & Mandryk, R. L. (2020). Effective gamification of the stop-signal task: Two controlled laboratory experiments. *JMIR Serious Games*, *8*(3), e17810. https://doi.org/10.2196/17810

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2. https://doi.org/10.1037/a0024338

Golubickis, M., Falbén, J. K., Ho, N. S., Sui, J., Cunningham, W. A., & Macrae, C. N. (2020). Parts of me: Identity-relevance moderates self-prioritization. *Consciousness and Cognition*, *77*, 102848. https://doi.org/10.1016/j.concog. 2019.102848

Golubickis, M., Falbén, J. K., Sahraie, A., Visokomogilski, A., Cunningham, W. A., Sui, J., & Macrae, C. N. (2017). Self-prioritization and perceptual matching: The effects of temporal construal. *Memory & Cognition*, *45*, 1223–1239. https://doi.org/10.3758/s13421-017-0722-3

*Golubickis, M., & Macrae, C. N. (2021). Judging me and you: Task design modulates self-prioritization. *Acta Psychologica*, *218*, 103350. https://doi.org/10.1016/j.actpsy.2021.103350

Green, S. B., Yang, Y., Alt, M., Brinkley, S., Gray, S., Hogan, T., & Cowan, N. (2016). Use of internal consistency coefficients for estimating reliability of experimental task scores. *Psychonomic Bulletin & Review*, *23*, 750–763. https: //doi.org/10.3758/s13423-015-0968-3

Groulx, J. T., Harding, B., & Cousineau, D. (2020). The ez diffusion model: An overview with derivation, software, and an application to the same-different task. *The Quantitative Methods for Psychology*, *16*(2), 154–174. https://doi. org/10.20982/tqmp.16.2.p154

*Haciahmet, C. C., Golubickis, M., Schäfer, S., Frings, C., & Pastötter, B. (2023). The oscillatory fingerprints of self-prioritization: Novel markers in spectral EEG for self-relevant processing. *Psychophysiology, 60*(12), e14396. https://doi.org/10.1111/psyp.14396

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2020). Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox. Preprint at *OSF* https://doi.org/10.31234/osf.io/xr7y3

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*, academic press.

*Hobbs, C., Sui, J., Kessler, D., Munafò, M. R., & Button, K. S. (2023). Self-processing in relation to emotion and reward processing in depression. *Psychological Medicine*, *53*(5), 1924–1936. https://doi.org/10.1017/S0033291721003597

*Hu, C.-P., Peng, K., & Sui, J. (2023). Data for training effect of self prioritization[ds/ol]. v2. *Science Data Bank*. https://doi.org/10.57760/sciencedb.08117

Hu, C.-P., Lan, Y., Macrae, C. N., & Sui, J. (2020). Good me bad me: Prioritization of the good-self during perceptual decision-making. *Collabra. Psychology*, *6*(1), 20. https://doi.org/10.1525/collabra.301

Hughes, M. M., Linck, J. A., Bowles, A. R., Koeth, J. T., & Bunting, M. F. (2014). Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity. *Behavior Research Methods*, *46*(3), 702–721. https://doi.org/10.3758/s13428-013-0411-5

Hughes, S. M., & Harrison, M. A. (2013). I like my voice better: Self-enhancement bias in perceptions of voice attractiveness. *Perception*, *42*(9), 941–949. https://doi.org/10.1068/p7526

Humphreys, G. W., & Sui, J. (2015). The salient self: Social saliency effects based on self-bias. *Journal of Cognitive Psychology*, *27*(2), 129–140. https://doi.org/10.1080/20445911.2014.996156

Kahveci, S. (2020). *AATtools: Reliability and scoring routines for the approach-avoidance task*.

Kahveci, S., Bathke, A., & Blechert, J. (2022). Reliability of reaction time tasks: How should it be computed? Preprint at *OSF* https://doi.org/10.31234/osf.io/ta59r

Karvelis, P., Paulus, M. P., & Diaconescu, A. O. (2023). Individual differences in computational psychiatry: A review of current challenges. *Neuroscience & Biobehavioral Reviews*, 105137. https://doi.org/10.1016/j.neubiorev.2023. 105137

Keenan, J. P., Wheeler, M. A., Gallup, G. G., & Pascual-Leone, A. (2000). Self recognition and the right prefrontal cortex. *Trends in Cognitive Sciences*, *4* (9), 338–344. https://doi.org/10.1016/S1364-6613(00)01521-7

Kircher, T. T., Senior, C., Phillips, M. L., Benson, P. J., Bullmore, E. T., Brammer, M., Simmons, A., Williams, S. C., Bartels, M., & David, A. S. (2000). Towards a functional neuroanatomy of self-processing: Effects of faces and words. *Cognitive Brain Research*, *10*(1-2), 133–144. https://doi.org/10.1016/S09266410(00)00036-7

Kline, P. (2015). *A handbook of test construction (psychology revivals): Introduction to psychometric design.* Routledge.

*Kolvoort, I. R., Wainio-Theberge, S., Wolff, A., & Northoff, G. (2020). Temporal integration as "common currency" of brain and self-scale-free activity in resting-state eeg correlates with temporal delay effects on self-relatedness. *Human Brain Mapping*, *41*(15), 4355–4374. https://doi.org/10.1002/hbm. 25129

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kucina, T., Wells, L., Lewis, I., de Salas, K., Kohl, A., Palmer, M. A., Sauer, J. D., Matzke, D., Aidman, E., & Heathcote, A. (2023). Calibration of cognitive tests to address the reliability paradox for decision-conflict tasks. *Nature Communications*, *14*(1), 2234. https://doi.org/10.1038/s41467-023-377772

Kupper, L. L., & Hafner, K. b. (1989). On assessing interrater agreement for multiple attribute responses. *Biometrics*, *45*(3), 957–967. https://doi.org/10.2307/ 2531695

*Liang, Q., Wang, Y., Wang, F., Li, Z., & Li, D. (2021). Prioritization of personally relevant stimuli in male abstinent heroin users. *Journal of Psychiatric Research, 142*, 132–139. https://doi.org/10.1016/j.jpsychires.2021.07.058

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation–a discussion and demonstration of basic features. *PloS One*, *14*(7), e0219854. https://doi.org/10.1371/journal.pone.0219854

Lin, H., Saunders, B., Friese, M., Evans, N. J., & Inzlicht, M. (2020). Strong effort manipulations reduce response caution: A preregistered reinvention of the egodepletion paradigm. *Psychological Science*, *31*(5), 531–547. https://doi.org/10.1177/0956797620904990

Liu, Z., Hu, C.-P, & Sui, J. (2021). Is Self-associative Learning Unique? A Meta-analytical comparison of the prioritization effect of self- and non-self associative learning. Pre-registration at *OSF* https://doi.org/10.17605/OSF.IO/EUQMF

Liu, Song, Y., Lee, N. A., Bennett, D. M., Button, K. S., Greenshaw, A., Cao, B., & Sui, J. (2022). Depression screening using a non-verbal self-association task: A machine-learning based pilot study. *Journal of Affective Disorders*, *310*, 87–95. https://doi.org/10.1016/j.jad.2022.04.122

*Liu, T., Sui, J. & Hildebrandt, A. To see or not to see: the parallel processing of self-relevance and facial expressions. (2023). *Cognitive Research: Principles and Implications*, *8*(1), 70. https://doi.org/10.1186/s41235-023-00524-8

Logie, R. H., Sala, S. D., Laiacona, M., Chalmers, P., & Wynn, V. (1996). Group aggregates and individual reliability: The case of verbal short-term memory. *Memory & Cognition*, *24*, 305–321. https://doi.org/10.3758/BF03213295

Macrae, C. N., Visokomogilski, A., Golubickis, M., Cunningham, W. A., & Sahraie, A. (2017). Self-relevance prioritizes access to visual awareness. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(3), 438. https://doi.org/10.1037/xhp0000361

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Martínez-Pérez, V., Campoy, G., Palmero, L. B., & Fuentes, L. J. (2020). Examining the dorsolateral and ventromedial prefrontal cortex involvement in the self-attention network: A randomized, sham-controlled, parallel group, doubleblind, and multichannel hd-tdcs study. *Frontiers in Neuroscience*, *14*, 683. https://doi.org/10.3389/fnins.2020.00683

*Martínez-Pérez, V., Sandoval-Lentisco, A., Tortajada, M., Palmero, L. B., Campoy, G., & Fuentes, L. J. (2024). Self-prioritization effect in the attentional blink paradigm: Attention-based or familiarity-based effect? *Consciousness and Cognition, 117*, 103607. https://doi.org/10.1016/j.concog.2023.103607

Moseley, R. L., Liu, C. H., Gregory, N. J., Smith, P., Baron-Cohen, S., & Sui, J. (2022). Levels of self-representation and their sociocognitive correlates in late-diagnosed autistic adults. *Journal of Autism and Developmental Disorders*, *52*(7), 3246–3259. https://doi.org/10.1007/s10803-021-05251-x

*Navon, M., & Makovski, T. (2021). Are self-related items unique? the self-prioritization effect revisited. Preprint at *OSF* https://doi.org/10.31234/osf.io/9dzm4

Nijhof, A. D., Shapiro, K. L., Catmur, C., & Bird, G. (2020). No evidence for a common self-bias across cognitive domains. *Cognition*, *197*, 104186. https://doi.org/10.1016/j.cognition.2020.104186

Orellana-Corrales, G., Matschke, C., & Wesslein, A.-K. (2020). Does self-associating a geometric shape immediately cause attentional prioritization? comparing familiar versus recently self-associated stimuli in the dot-probe task. *Experimental Psychology*, *67*(6), 335. https://doi.org/10.1027/1618-3169/ a000502

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, *88*, 105906. https://doi.org/10. 1136/bmj.n71

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

Payne, B., Lavan, N., Knight, S., & McGettigan, C. (2021). Perceptual prioritization of self-associated voices. *British Journal of Psychology*, *112*(3), 585–610. https://doi.org/10.1111/bjop.12479

*Perrykkad, K., & Hohwy, J. (2022). How selves differ within and across cognitive domains: self-prioritisation, self-concept, and psychiatric traits. *BMC Psychology*, *10*(1), 165. https://doi.org/10.1111/10.1186/s40359-022-00870-0

Pronk, T., Hirst, R. J., Wiers, R. W., & Murre, J. M. (2023). Can we measure individual differences in cognitive measures reliably via smartphones? A comparison of the flanker effect across device types and samples. *Behavior Research Methods*, *55*(4), 1641–1652. https://doi.org/10.3758/s13428-022-01885-6

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment.

Psychonomic Bulletin & Review, *29*(1), 44–54. https: //doi.org/10.3758/s13423-021-01948-3

*Qian, H., Wang, Z., Li, C., & Gao, X. (2020). Prioritised self-referential processing is modulated by emotional arousal. *Quarterly Journal of Experimental Psychology*, *73*(5), 688–697. https://doi.org/10.1177/1747021819892158

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics*, 201–211.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.Rproject.org/

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115

Reuther, J., & Chakravarthi, R. (2017). Does self-prioritization affect perceptual processes? *Visual Cognition*, *25*(1-3), 381–398. https://doi.org/10.1080/ 13506285.2017.1323813

Revelle, W. R. (2017). Psych: Procedures for personality and psychological research. https://CRAN.Rpro ject.org/package=psych

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology*, *35*(9), 677–88. https://doi.org/ 10.1037//0022-3514.35.9.677

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. https: //doi.org/10.3758/s13423-018-1558-y

*Schäfer, S., & Frings, C. (2019). Understanding self-prioritisation: The prioritisation of self-relevant stimuli and its relation to the individual self-esteem. *Journal of Cognitive Psychology*, *31*(8), 813–824. https://doi.org/10.1080/20445911. 2019.1686393

Scheller, M., & Sui, J. (2022a). The power of the self: Anchoring information processing across contexts. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(9), 1001–1021. https://doi.org/10.1037/xhp0001017

Scheller, M., & Sui, J. (2022). Social relevance modulates multisensory integration. *Journal of Experimental Psychology. Human Perception and Performance*, *48*(9), 1022–1038. https://doi.org/10.1037/xhp0001013

Shieh, G. (2010). Estimation of the simple correlation coefficient. *Behavior Research Methods*, *42*(4), 906–917. https://doi.org/10.3758/BRM.42.4.906

Smithson, C. J., Chow, J. K., Chang, T. Y., & Gauthier, I. (2024). Measuring object recognition ability: Reliability, validity, and the aggregate z-score approach. *Behavior Research Methods*, 1–15. https://doi.org/10.3758/s13428-024-02372-w

Stoeber, J., & Eysenck, M. W. (2008). Perfectionism and efficiency: Accuracy, response bias, and invested time in proof-reading performance. *Journal of Research in Personality*, *42*(6), 1673–1678. https://doi.org/10.1016/j.jrp.2008.08.001

Sui, J., Cao, B., Song, Y., & Greenshaw, A. J. (2023). Individual differences in self-and value-based reward processing. *Current Research in Behavioral Sciences*, 4, 100095. https://doi.org/10.1016/j.crbeha.2022.100095

*Sui, J., He, X., Golubickis, M., Svensson, S. L., & Macrae, C. N. (2023). Electrophysiological correlates of self-prioritization. *Consciousness and Cognition, 108*, 103475. https://doi.org/10.1016/j.concog.2023.103475

Sui, J., He, X., & Humphreys, G. W. (2012). Perceptual effects of social salience: Evidence from self-prioritization effects on perceptual matching. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1105–1117. https://doi.org/10.1037/a0029792

Sui, J., & Humphreys, G. W. (2017). The self survives extinction: Self-association biases attention in patients with visual extinction. *Cortex*, *95*, 248–256. https://doi.org/10.1016/j.cortex.2017.08.006

*Sui, J., Sun, Y., Peng, K., & Humphreys, G. W. (2014). The automatic and the expected self: Separating self-and familiarity biases effects by manipulating stimulus probability. *Attention, Perception, & Psychophysics, 76*, 1176–1184. https://doi.org/10.3758/s13414-014-0631-5

*Svensson, S. L., Golubickis, M., Maclean, H., Falb´en, J. K., Persson, L. M., Tsamadi, D., Caughey, S., Sahraie, A., & Macrae, C. N. (2022). More or less of me and you: Self-relevance augments the effects of item probability on stimulus prioritization. *Psychological Research*, *86*(4), 1145–1164. https://doi.org/10. 1007/s00426-021-01562-x

Turk, D. J., Heatherton, T. F., Kelley, W. M., Funnell, M. G., Gazzaniga, M. S., & Macrae, C. N. (2002). Mike or me? self-recognition in a split-brain patient. *Nature Neuroscience*, *5*(9), 841–842. https://doi.org/10.1038/nn907

Vermeiren, A., & Cleeremans, A. (2012). The validity of d' measures. *PloS One*, *7*(2), e31595. https://doi.org/10.1371/journal.pone.0031595

*Vicovaro, M., Dalmaso, M., & Bertamini, M. (2022). Towards the boundaries of self-prioritization: Associating the self with asymmetric shapes disrupts the self-prioritization effect. *Journal of Experimental Psychology: Human Perception and Performance, 48*(9), 972. https://doi.org/10.1037/xhp0001036

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036. i03

Wabersich, D., & Vandekerckhove, J. (2014). The rwiener package: An r package pro-viding distribution functions for the wiener diffusion model. *R Journal*, *6*(1). https://doi.org/10.32614/RJ-2014-005

Wagenmakers, E.-J., Van Der Maas, H. L., & Grasman, R. P. (2007). An ez-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14* (1), 3–22. https://doi.org/10.3758/BF03194023

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modelling of behavioral data. *Elife*, *8*, e49547. https://doi.org/10.7554/eLife. 49547

*Woźniak, M., Kourtis, D., & Knoblich, G. (2018). Prioritization of arbitrary faces associated to self: An eeg study. *PloS One*, *13*(1), e0190679. https://doi.org/10.1371/journal.pone.0190679

*Woźniak, M., & Knoblich, G. (2022). Self-prioritization depends on assumed task-relevance of self-association. *Psychological Research, 86*(5), 1599–1614. https://doi.org/10.1007/s00426-021-01584-5

Xu, Kiar, G., Cho, J. W., Bridgeford, E. W., Nikolaidis, A., Vogelstein, J. T., & Milham, M. P. (2023). Rex: An integrative tool for quantifying and optimizing measurement reliability for the study of individual differences. *Nature Methods*, 1–4. https://doi.org/10.1038/s41592-023-01901-3

*Xu, Yuan, Y., Xie, X., Tan, H., & Guan, L. (2021). Romantic feedbacks influence self-relevant processing: The moderating effects of sex difference and facial attractiveness. *Current Psychology*, 1–13. https://doi.org/10.1007/s12144021-02114-7

Yankouskaya, A., Lovett, G., & Sui, J. (2023). The relationship between self, value-based reward, and emotion prioritisation effects. *Quarterly Journal of Experimental Psychology*, *76*(4), 942-960. https://doi.org/10.1177/174702182211028

Yue, L., Zuo, X.-N., & Chuan-Peng, H. (2023). The weird problem in a "non-weird" context: A meta-research on the representativeness of human subjects in Chinese psychological research. Pre-registration at *OSF* https://doi.org/osf.io/y9hwq

Zhang, Y., Wang, F., & Sui, J. (2023). Decoding individual differences in self-prioritization from the resting-state functional connectome. *NeuroImage*, 120205. https://doi.org/10.1016/j.neuroimage.2023.120205

Zorowitz, S., & Niv, Y. (2023). Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging.* https://doi.org/10.1016/j.bpsc.2023.02.004

Supplementary Material for "A Multiverse Assessment of the Reliability of the Self Matching Task as a Measurement of the Self-Prioritization Effect"

Zheng Liu[1,2†], Mengzhen Hu[1†], Yuanrui Zheng[1], Jie Sui[3], Hu Chuan-Peng[1*]

[1*]School of Psychology, Nanjing Normal University, Nanjing, China.

[2*]School of Humanities and Social Science, The Chinese University of Hong Kong-Shenzhen, Shenzhen, China.

[3*]School of Psychology, University of Aberdeen, Old Aberdeen, Scotland.

[*]Corresponding author(s). E-mail(s): hu.chuan-peng@nnu.edu.cn; hcp4715@hotmail.com;
[†]These authors contributed equally to this work.

1 Supplementary Methods

1.1 Methodological details of the dataset from Hu et al. (2023)

In this current study, we utilized a dataset that was previously collected by our research team in 2016 (Hu et al., 2023). The original study aimed to compare SPE between two groups: individuals with sub-clinical depression and those without depression. The dataset comprised data from six time points, each one week apart, collected from a sample of 36 participants recruited from the Tsinghua University community. At each time point, participants completed three distinct tasks: Experiment A (a modified SMT with large deviations), Experiment B (another modified SMT with small deviations), and a questionnaire. The original research faced challenges in recruiting sub-clinical depressed participants, leading to an overrepresentation of individuals in the healthy control group, however, making it suitable for the current study. Thus, in our current analysis, we focused on the subset of data related to the neutral condition in Experiment B from these 36 participants. In the following sections, we provided a detailed overview of the original experimental design.

1.1.1 Ethics Information

The experiment was approved by the IRB at the Department of Psychology, Tsinghua University, and all participants provided informed consent.

1.1.2 Participants.

36 participants were recruited from Tsinghua University and the nearby community, all of whom were right-handed and had normal or corrected-to-normal vision. Participants were pre-tested for their depressive level by Beck Depression Inventory-II (BDI-II) (Wang et al., 2011). Data from three participants were excluded due to invalid trials or program malfunctions. The exclusion left 33 valid participants ($Mean_{age} = 21.06$, $SD_{age} = 3.24$), with 21 females and 12 males. It's worth noting that within this sample of 33 participants, only six individuals had a BDI-II score exceeding 20.

1.1.3 Experimental Design

Experiment 2 was a 2 (Matching: Matching vs. Non-matching) × 3 (Identity: Self, Friend, Stranger) × 4 (Emotion: Control, Neutral, Happy, Sad) × 6 (Sessions: 1-6) experiment.

1.1.4 Procedure

The experiment was finished individually in a dimly lighted room. Stimuli were presented and responses were collected using E-Prime 2.0 on PC. The monitor was at 1024 × 768 resolution with a 100 Hz refresh rate.
The experiment has two phases (see Fig. S1). Following Sui et al. (2012), the first phase comprised an instructional stage in which participants were required to associate geometric shapes with labels. The instruction stage lasted for approximately 60 seconds and shape-target associations were counterbalanced across the sample. Next, participants performed a matching

task. At the start of each trial, a fixation cross was first displayed in the center of the screen for 500 ms. Then, a shape–label pairing as well as the fixation cross were presented for 100ms, respectively. The next frame showed a blank screen for 800-1200 ms. Participants were asked to determine whether the shape was appropriately matched to the label by pressing one of the two response buttons as quickly and precisely as possible within this timeframe.

The participants needed to separately learn 4 sets of associations between shapes and labels. The associations contained 1 control condition and 3 sets of emotion-based conditions. In the control condition, participants learned the association between 3 geometric shapes (circle, horizontal ellipse and vertical ellipse) and three labels (Self, Friend, Stranger). In each of the emotion-based conditions, participants would see facial expressions (happy, sad, neutral) appear on the circle, horizontal ellipse and vertical ellipse (see Fig. S1). In each condition, before commencing the formal experimental trials, participants underwent a training session comprising 24 practice trials. After the practice trials, each participant completed 6 blocks of 60 trials in the task. There were six types of shape-label associations: Matching (Matching / Non-matching) x Shape (Self, Friend, Stranger) associations, with 60 trials for each association. Participants took a short break (up to 60 seconds) after each block. Each participant was required to repeat the experiment six times, with a one-week gap between each wave of experiments.

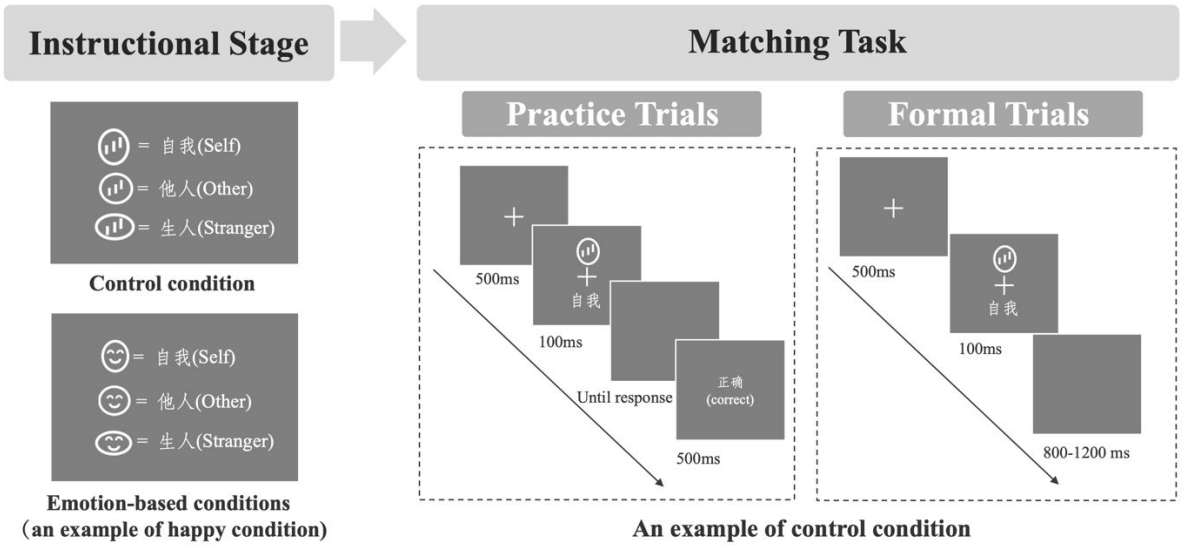


Fig. S1 Procedure of the SMT in Experiment B (Hu et al., 2023). *Note:* The labels and feedback appeared in Chinese in the experiment. In the associative learning task, the matched associations of shapes and labels were counterbalanced between participants. Timely feedback was not provided in formal trials.

1.2 Paper Selection Procedure

In Figure S2, we presented the detailed paper selection procedure.

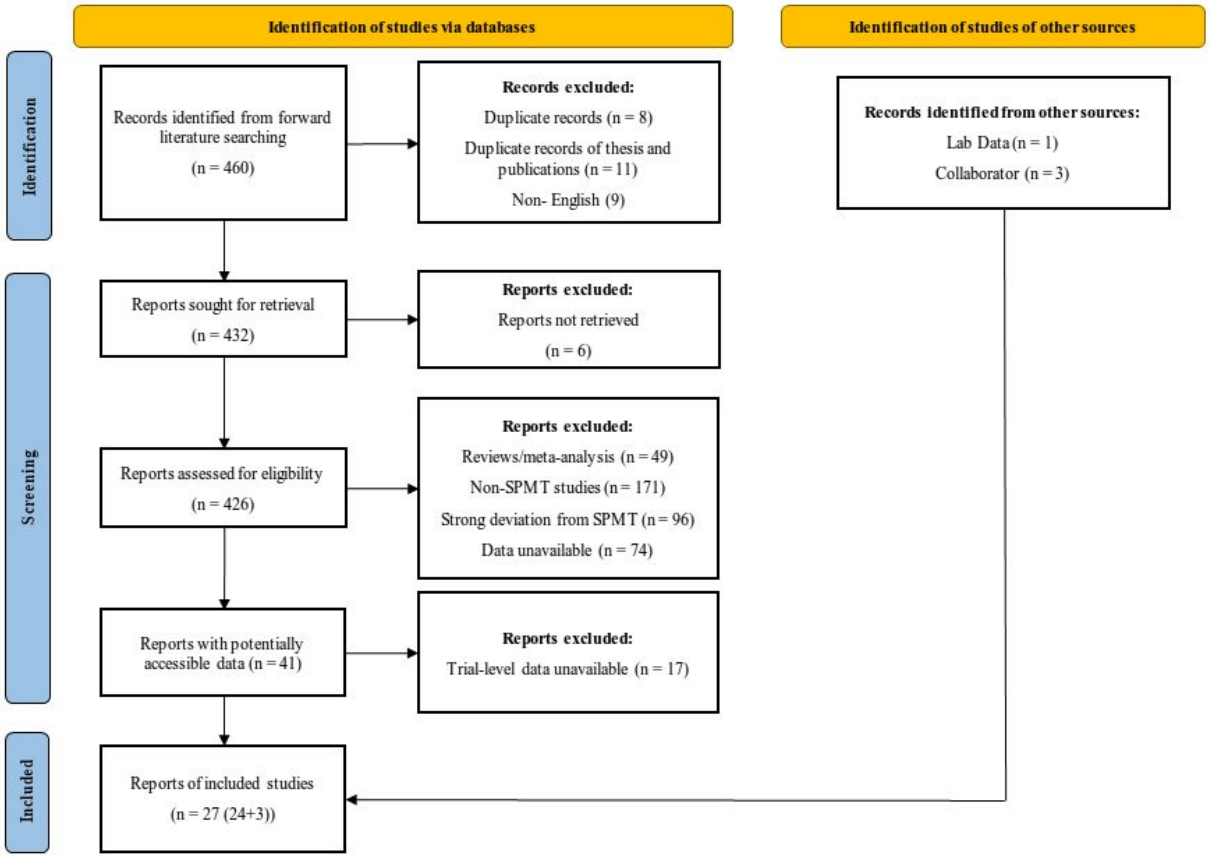


Figure S2. Paper Selection Procedure (adapted from PRISMA Flow Diagram (Page et al., 2021)).

1.3 Parameter Recovery Result for Package Comparison

We chose not to utilize the HDDM package (Wiecki et al., 2013) since the computation process was significantly time-consuming, necessitating high computational resources and leading to prolonged overall analysis time. Instead, we performed a package comparison by generating 100 datasets using the HDDM package in Python, in order to identify the most appropriate package for our analysis. These datasets were specifically configured with parameters $a = 2$, $t = 0.3$, $v = 1$, and $z = 0.7$.

Subsequently, we utilized three widely used DDM packages in R, namely RWiener (Viechtbauer, 2010), hausekeep (Lin, 2019), and FastDMinR (Voss & Voss, 2007), to compute parameter estimates for these generated datasets. The evaluation process involved comparing the computed values obtained from the R packages with the set parameters. If the computed values from the R packages were found to be closer to the set values, it signified that the respective R package provided more accurate parameter estimation for the DDM.

Fig. S3 presents the results of the package comparison. The estimated drift rate (v) obtained from RWiener was 1.01, with a 95% confidence interval of [.98, 1.03], which is closely aligned with our pre-defined values. Similarly, the estimated starting point (z) is 0.77, with a 95% confidence interval of [.76, .78], also very close to our pre-defined value. On the contrary, the parameters

calculated using other packages either showed high inaccuracies, excessively wide confidence intervals or required extended computation times. As a result, we have opted to utilize RWiener for our calculations. It struck a favourable balance between accuracy, confidence interval width, and computational efficiency, making it the most suitable choice for our analysis.

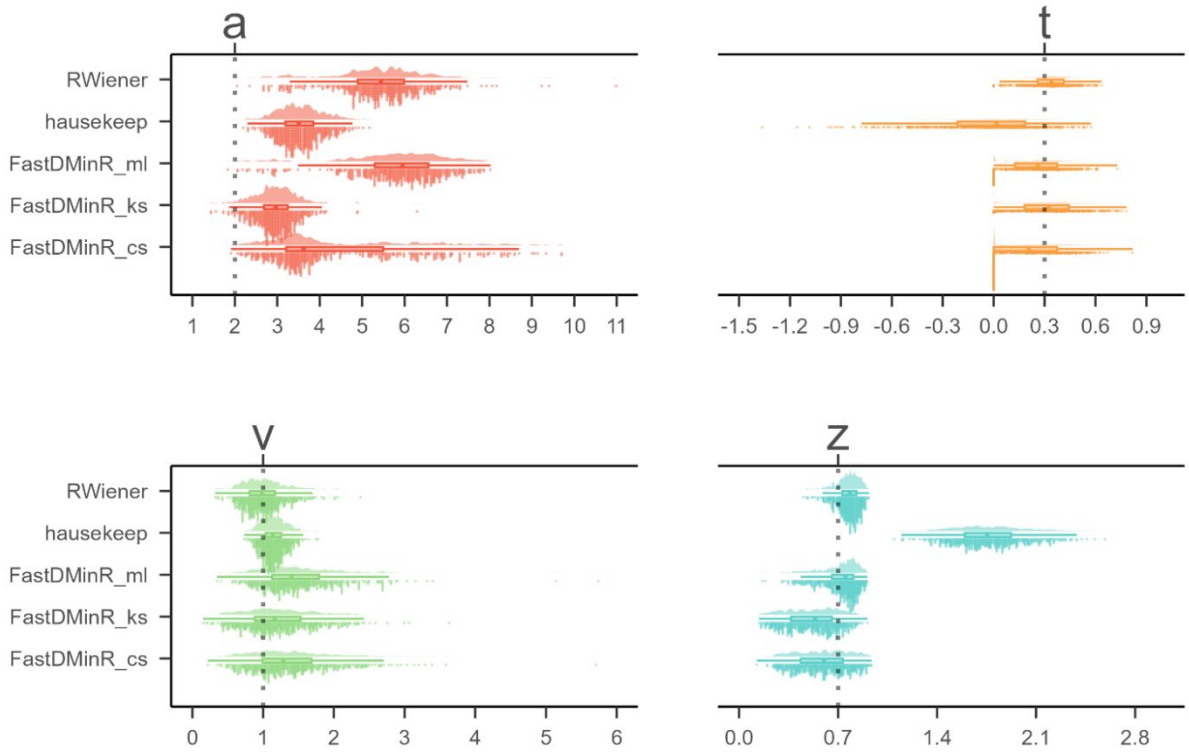


Fig. S3 DDM Packages Comparison. *Note:* The parameters of interest in the Drift-Diffusion Model (DDM) are represented as follows: "*a*" denotes the threshold parameter, "*t*" represents the non-decision time, "*v*" indicates the drift rate, and "*z*" corresponds to the starting point. The y-axis of the graph displays the estimation of these DDM parameters using three different R packages: "RWiener," "hausekeep," and "FastDMinR." In total, there are five methods for estimating DDM parameters, with three methods originating from the "FastDMinR" package. On the x-axis, the values of the estimated parameters are plotted. The dashed line on the graph indicates the true value of the parameter being estimated.

2 Supplementary Results

2.1 Group Level SPE for Other Measures

We conducted a meta-analysis of all the 6 indicators of SPE. The forest plots are presented in Fig. S3.

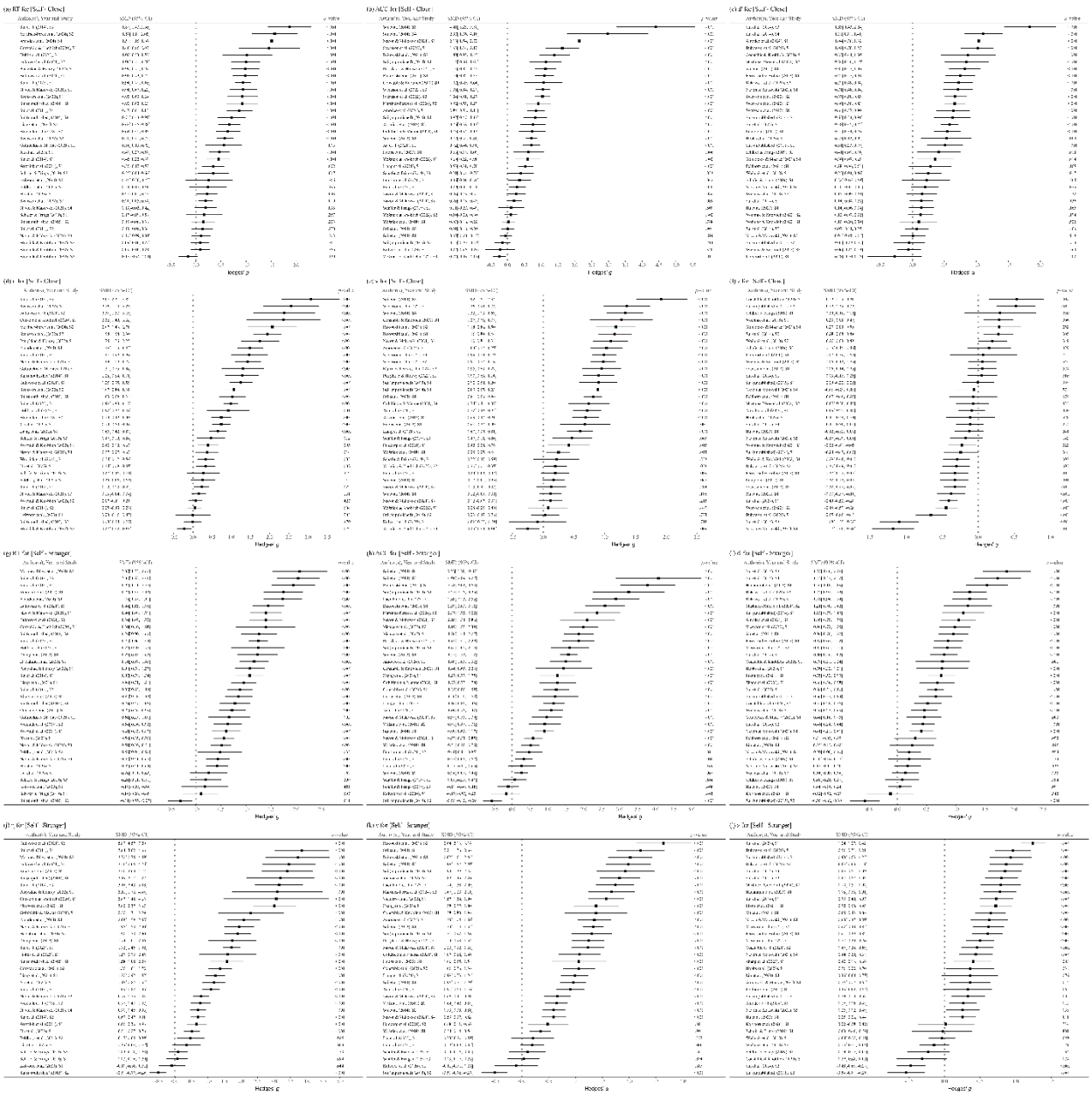


Fig. S4 (a) Forest Plot for SPE Measures. *Note*: Fig (a)-(f) represent the forest plots corresponding to RT, ACC, *d'*, *η*, *v*, and *z* under the condition where Target is Close. Fig (g)-(l) represents the forest plots corresponding to *d'*, *η*, *v*, and *z* under the condition where Target is Stranger.

Due to the limited availability of papers on "Celebrity" and "Nonperson", we were unable to perform a meta-analysis on these baselines. Instead, we conducted paired-sample t-tests comparing self and baseline conditions. Hedges' *g* was calculated, and the results were presented in Table. S1. Considering there is only one paper available for these baselines, it is advisable to approach these results with caution.

Table S1 T-test Results of SPE Measures in SMT

Baseline	Indicators	Hedges' *g* [95% CI]	# of Studies	Q	*p*	I^2
Celebrity						
	RT	1.76 [1.11, 2.41]	1			
	ACC	2.08 [1.39, 2.77]	1			
	d	1.41 [0.79, 2.03]	1			
	η	2.7 [1.93, 3.46]	1			
	v	1.46 [.83, 2.08]	1			
	z	.01 [-.54, .56]	1			
Nonperson						
	RT	.61 [.35, .86]	4	8.79	.032	71.30%
	ACC	1.53 [.23, 2.83]	4	65.50	<.001	96.44%
	d	0.67 [.27, 1.08]	4	20.24	<.001	88.70%
	η	1.49 [.32, 2.65]	4	52.11	<.001	94.76%
	v	.63 [.33, .94]	4	15.55	.001	82.98%
	z	-.09 [-.19, .01]	4	4.61	.203	0.03%

2.2 Split-Half Reliability Using Three Splitting Approaches

In this section, we presented the Split-Half Reliability (SHR) results for the SPE measures using three split-half methods: first-second, odd-even and permutated. We also included the drift rate (*v*) and starting point (*z*) estimated from the "hausekeep" package in the analysis. However, it's important to highlight that the estimation of parameter "*a*" in "hausekeep" significantly deviates from the HDDM approach, primarily because of its assumption that $z = a / 2$ (refer to Fig. S2). As a result, we have chosen not to include the results obtained from this package in the main text. Nevertheless, we presented them here for reference and transparency. Please refer to Fig. S5 for the visual representation of the results.

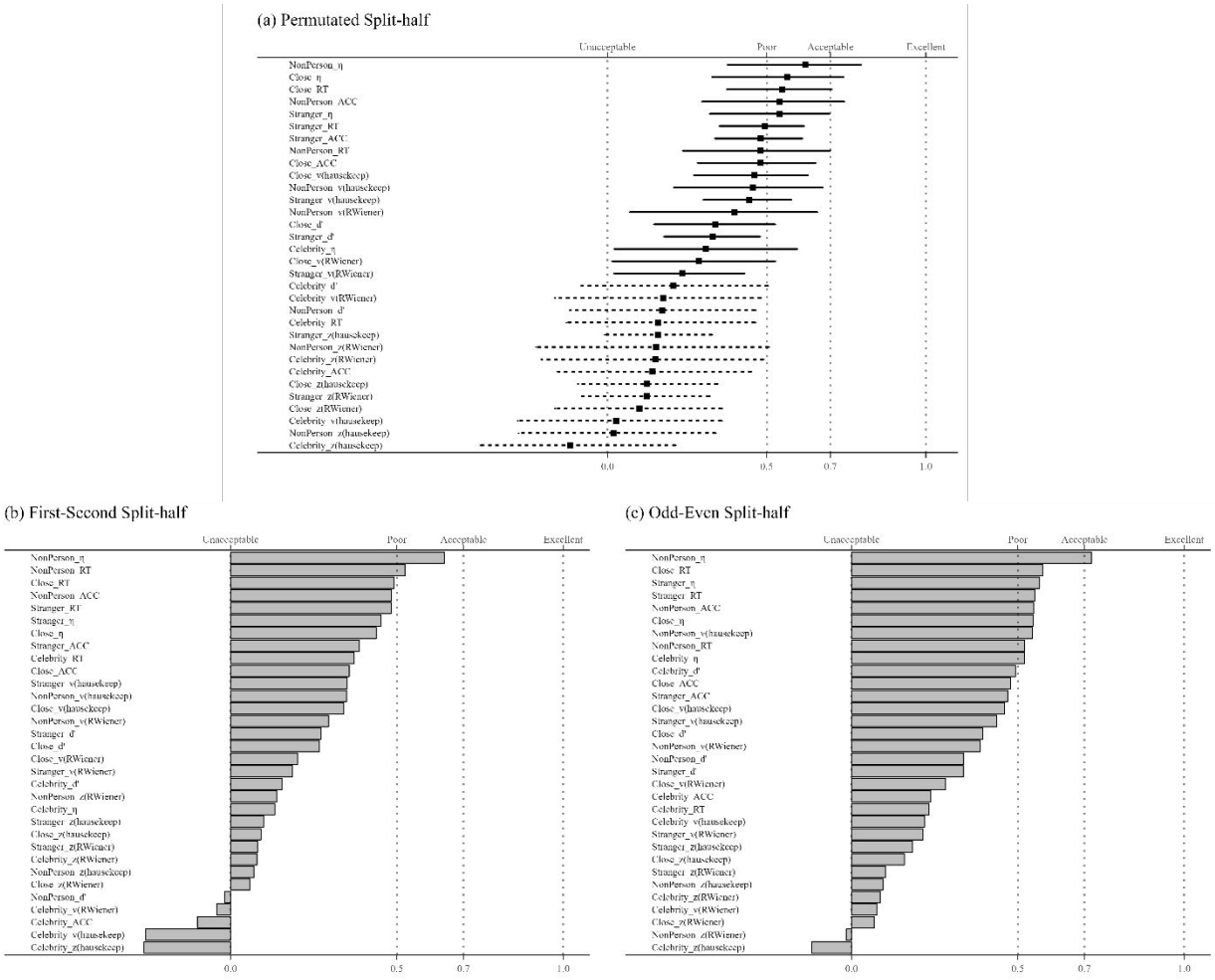


Fig. S5 Results of SHR Using Three Split-half Methods. (a) Results of SHR using Permutated Split-half Methods; (b) Results of SHR using First-Second Split-half Methods; (c) Results of SHR using Odd-Even Split-half Methods. *Note*: The vertical axis of the graph listed 32 different SPE measures, combining six indicators (RT, ACC, *d'*, *η*, *v*, *z*) and four baseline conditions (close other, stranger, celebrity, and non-person). The v and z implemented using the "hausekeep" package were also included. The weighted average split-half reliability and 95% confidence intervals are shown by points and lines. The figure is divided into separate facets arranged from left to right, each representing weighted average split-half reliability calculated using three distinct methods: first-second, odd-even and permutated.

The pattern of the results from the first-second split-half methods was similar to the permutated split-half method's outcomes. The top four split-half reliabilities, ranked highest, were as follows: Reaction Time (RT) with the "Stranger" contrast, Efficiency (*η*) with the "Stranger" contrast, RT with the "Close other" contrast, *η* with the "Self vs Close" contrast. However, the results obtained from the odd-even split-half method were notably different from the other two methods. We hypothesize that this discrepancy may be attributed to the odd-even method's sensitivity to temporal dependencies, which could have been influenced by the inherent sequential nature of responses in the SMT. Further investigation into the presence and impact of

serial dependency in the data would be valuable to better understand the observed variations in the split-half reliabilities among the different methods.

2.3 ICCs for SPE Measures Using Another Dataset

In Fig. S6, we presented the results of the Intraclass Correlation Coefficients (ICC2) for the SPE measures, where drift rate (*v*) and starting point (*z*) estimated from the "hausekeep" package were also included. In Fig. S5(b), we extended our exploration of ICC2 to include the SPE measures from one additional dataset. However, the SMT used in this dataset deviated quite strongly from the original SMT paradigm. Due to these significant differences, ICC2 obtained from this dataset may reflect variations introduced by the modified SMT rather than directly comparable results to the original paradigm.

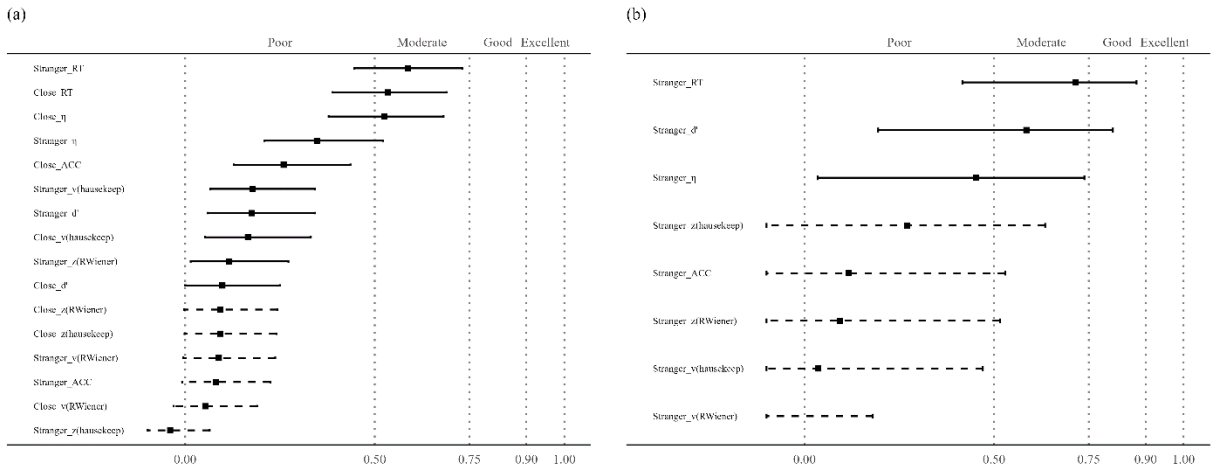


Fig. S6 ICCs for SPE Measures Using Hu et al. (2023) and Another Dataset. (a) ICC2 for SPE measures using Hu et al. (2023); (b) ICC2 for SPE measures using an additional dataset. *Note*: The vertical axis of the graph illustrates eight distinct indicators, which include two additional indices from the DDM, implemented using the "hausekeep" package. The line and dots on the graph represent the value of ICC2, along with their corresponding 95% confidence intervals. The dashed line indicates that the confidence interval for that point estimate extends beyond the range of our coordinate axes (0, 1).

Since the original design of Hu et al. (2023) incorporated measures from the Beck Depression Inventory-II (BDI-II) (Wang et al., 2011). Thus, in Fig. S7, we incorporated the BDI-II scores of individual participants as covariates when calculating ICC2. Notably, even after accounting for these BDI scores as covariates, we observed consistent ICC2 values both before and after this adjustment.

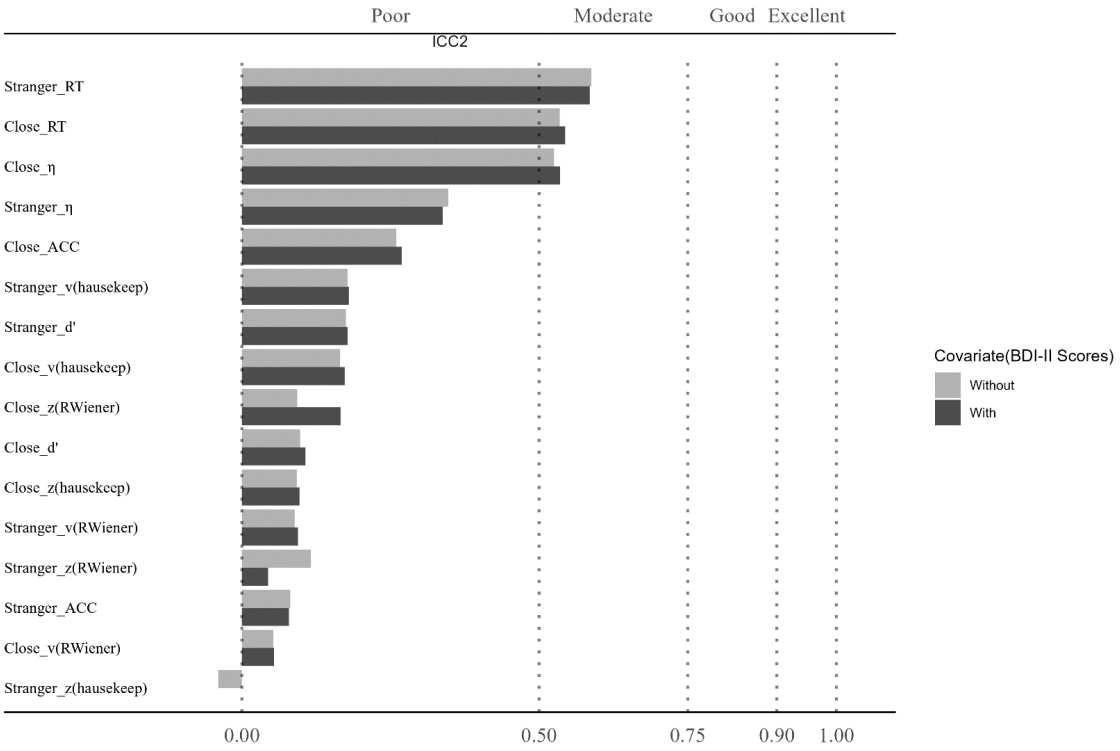


Fig. S7 ICC2 for SPE Measures Using Hu et al. (2023) with Covariant (BDI-II Scores). Note: The vertical axis of the graph illustrates eight distinct indicators, which include two additional indices from the DDM, implemented using the "hausekeep" package. The bar on the graph represents the value of ICC2.

2.4 Exploratory Analysis

In this section, we presented the results of the exploratory analysis of the current study. Our focus was on performing a correlation analysis that assessed the relationship between the number of trials and two key factors: permutated split-half reliability and effect size (Hedges' g). We also examine the relationship between permutated split-half reliability and effect size (Hedges' g). Furthermore, we adopted the Spearman-Brown prediction formula based on our current data to predict the trial counts at which the SMT achieves sufficient reliability.

We found significant correlations between trial numbers and permutated split-half reliability for some indicators, such as Reaction Time and Efficiency (see Fig. S8). However, for indicators like *d'* and *v*, the correlation with trial numbers was relatively weak.

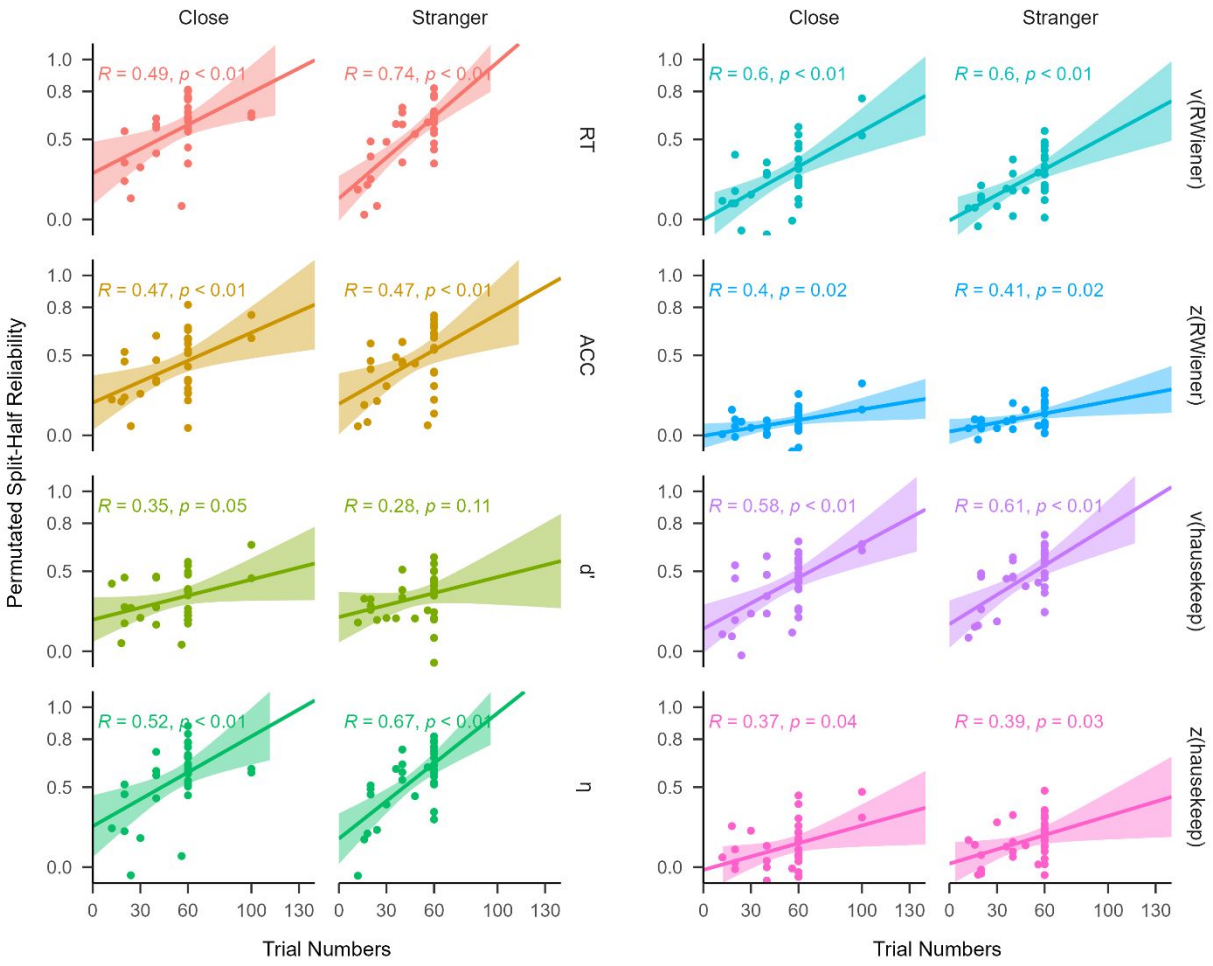


Fig. S8 Regression Analysis Between Permutated SHR and Trial Numbers Using Different SPE Measures. Note: The vertical axis represents the permutated split-half reliability, and the horizontal axis represents the number of trials. Each facet represents one SPE measure.

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.

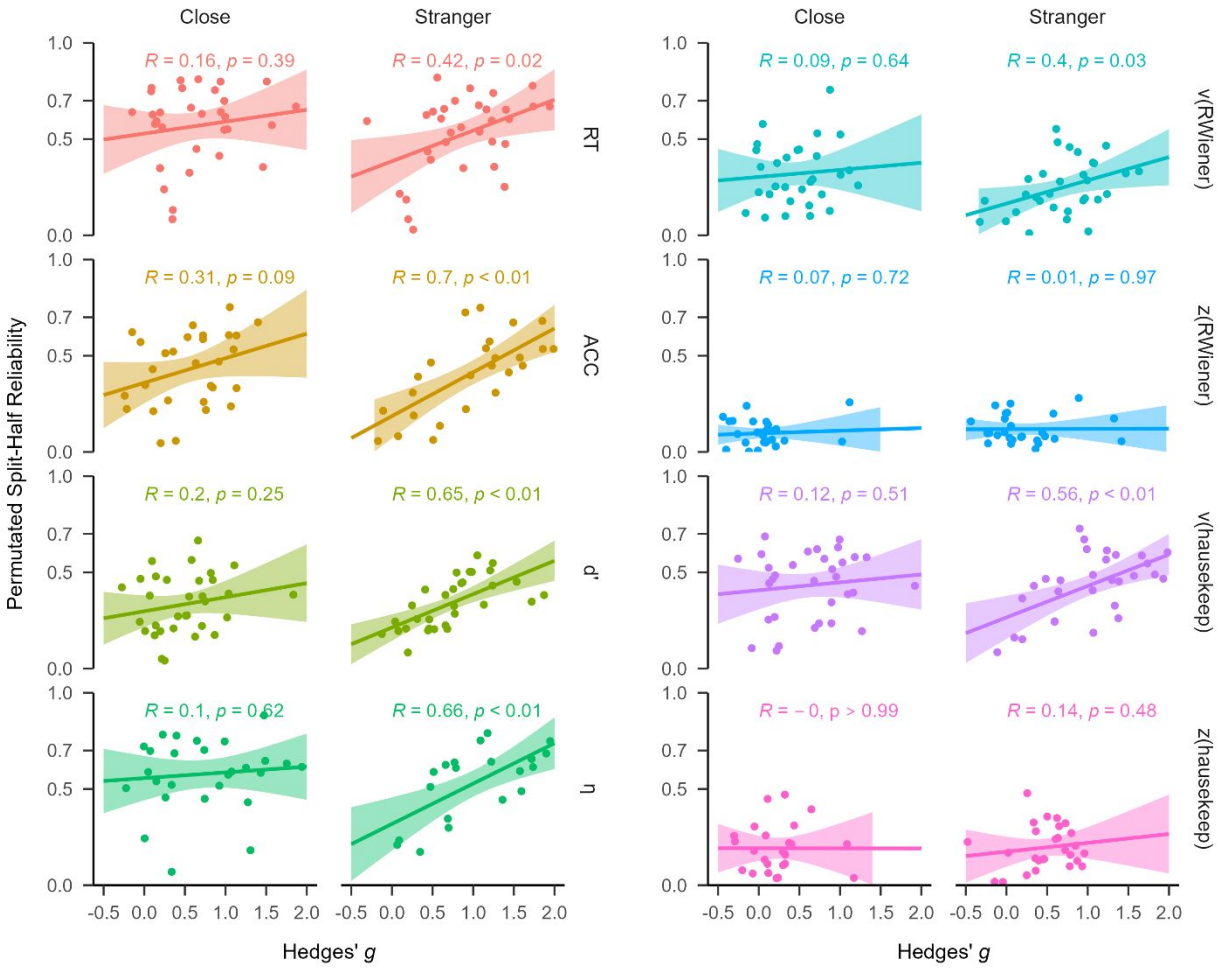


Fig. S9 Regression Analysis Between Permutated SHR and Effect Size (Hedges' g) Using Different SPE Measures. *Note*: The vertical axis represents permutated split-half reliability, and the horizontal axis represents the effect size (Hedges' *g*). Each facet represents one SPE measure.

We then calculated the correlation coefficient between trial numbers and effect size (Hedges' g), as shown in Fig. S10. No significant correlation was found.

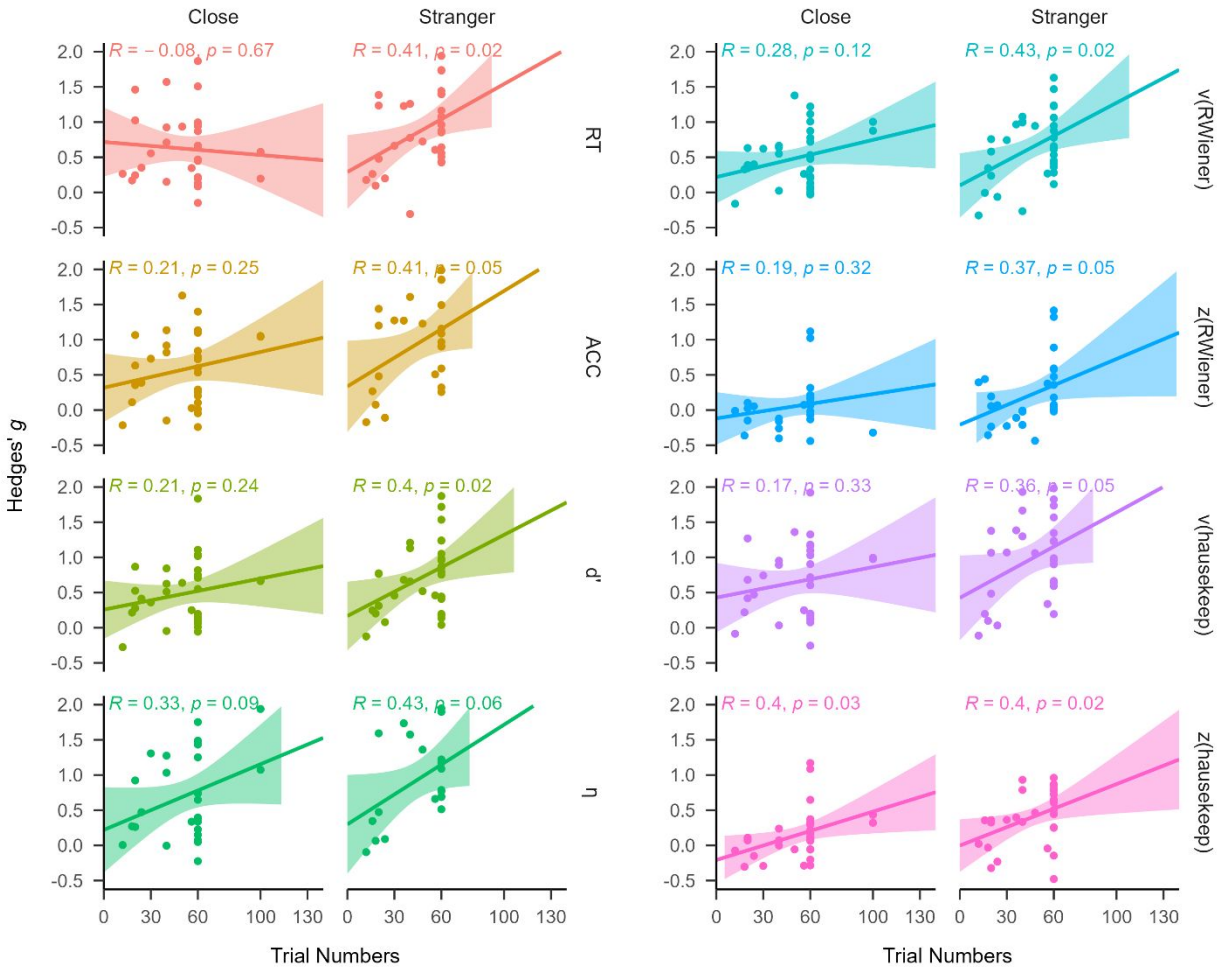


Fig. S10 Regression Analysis Between Trial Numbers and Effect Size (Hedges' g) Using Different SPE Measures. Note: The vertical axis represents the effect size (Hedges' g), and the horizontal axis represents trial numbers. Each facet represents one SPE measure.

Finally, we used the Spearman-Brown prediction formula to predict the trial numbers required for different levels of reliability. The results indicated that the number of trials required for achieving 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. For other SPE indices, more trials are required (see Fig. S11). It's important to emphasize that these findings are based on our current dataset and should be interpreted with caution. Further validation and verification of this relationship would be essential and will require new data collection efforts in future research.

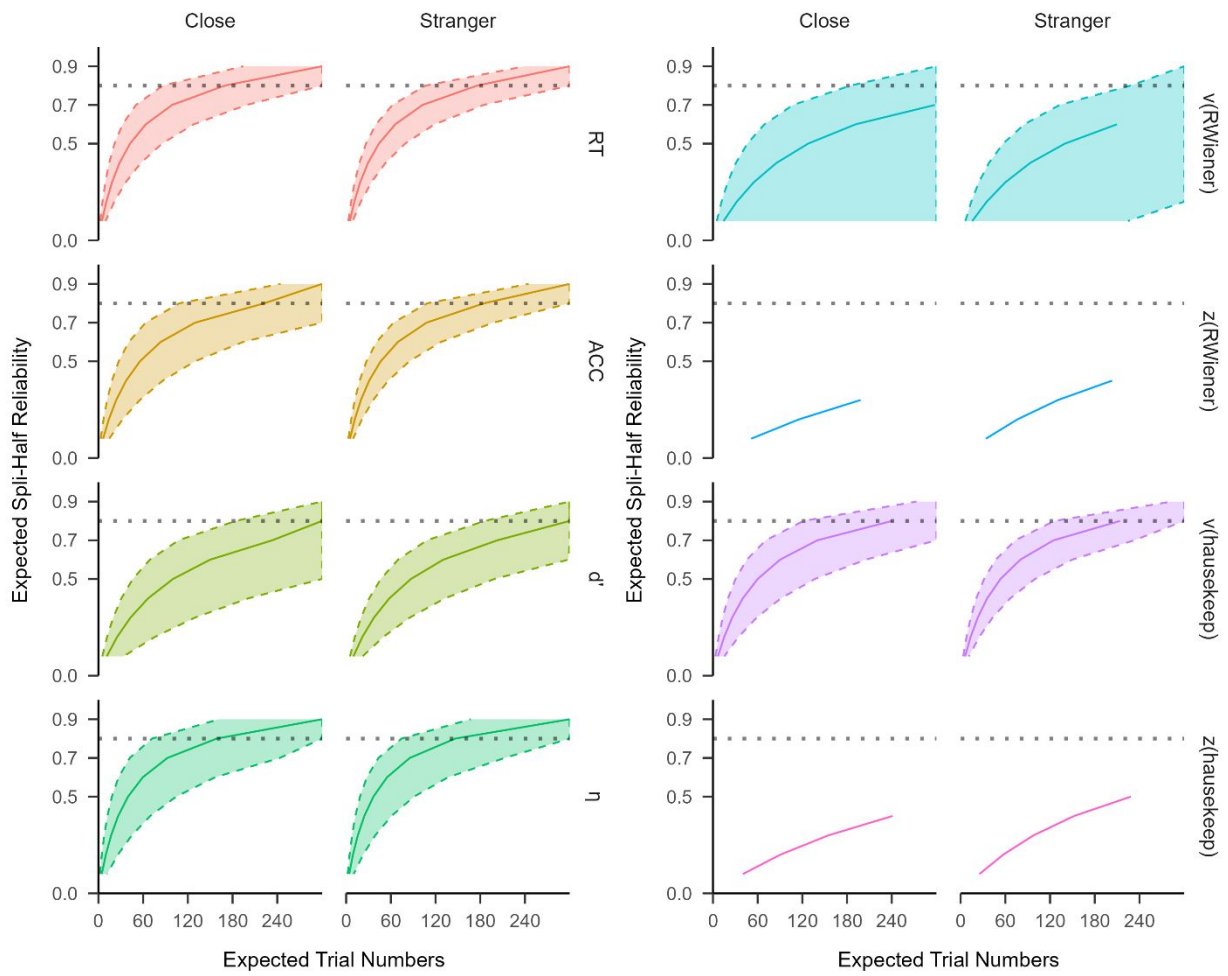


Fig. S11 Expected Trial Numbers Using Different SPE Measures. *Note*: The vertical axis represents the expected trial numbers calculated based on the spearman-brown function, and the horizontal axis represents the expected split-half reliability. Each facet represents one SPE measure. For SPE measured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

It's important to emphasize that the exploratory analysis was not part of the pre-registered plan, and our primary aim was not to provide a well-validated improvement for SMT. Further validation and verification of this relationship would be essential and will require new data collection efforts in future research. Nevertheless, taking into account the noteworthy correlation observed between the number of trials and permutated split-half reliability, our results indicated that when employing the SMT paradigm for individual differences, achieving higher reliability would likely require an increase in the number of conducted trials.

References

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hu, C.-P., Peng, K., & Sui, J. (2023). Data for training effect of self-prioritization[ds/ol]. v2. *Science Data Bank*. https://doi.org/10.57760/sciencedb. 08117

Lin, H. (2019). How to use hausekeep. https://doi.org/10.5281/zenodo.2555874

Logie, R. H., Sala, S. D., Laiacona, M., Chalmers, P., & Wynn, V. (1996). Group aggregates and individual reliability: The case of verbal short-term memory. *Memory & Cognition*, *24*, 305–321. https://doi.org/10.3758/BF03213295

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology*, *134*, 178–189. https://doi.org/10.1136/bmj.n71

Sui, J., He, X., & Humphreys, G. W. (2012). Perceptual effects of social salience: Evidence from self-prioritization effects on perceptual matching. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1105– 1117. https://doi.org/10.1037/a0029792

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036. i03

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*(4), 767–775. https://doi.org//10. 3758/BF03192967

Wang, Z., Yuan, C.-M., Huang, J., Li, Z.-Z., Chen, J., Zhang, H.-Y., Fang, Y.-R., & Xiao, Z.-P. (2011). Reliability and validity of the chinese version of beck depression inventory-ii among depression patients. *Chinese Mental Health Journal.*

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). Hddm: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics*, *7*, 14– 14. https://doi.org/10.3389/fninf.2013.00014