BEHAVIOR RESEARCH METHODS

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

Journal:	*Behavior Research Methods*
Manuscript ID	Draft
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	n/a
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

A Multiverse Assessment of the Reliability of the Perceptual 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 Sciences, 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 Perceptual Matching Task (SPMT) is a widely used task to investigate the cognitive processes underlying the Self-Prioritization Effect (SPE), wherein performance is enhanced for self-associated stimuli compared to other-associated ones. Despite the wide use of SPMT, there is a lack of attention on its reliability assessment. This ignorance is concerning, given the prevalence of the reliability paradox in cognitive tasks: cognitive tasks demonstrate relatively low reliability when evaluating individual differences, though they produce robust experimental effects. To fill this gap, this preregistered study investigated the reliability of SPMT using a multiverse approach, combining all possible indicators and baselines used to quantify SPE in SPMT. We examined the robustness and the reliability of 24 SPE measures across 17 datasets (N = 805). More specifically, we used a meta-analytical approach to estimate the robustness of SPE across datasets. We calculated the Split-Half Reliability (r) and Intraclass Correlation Coefficient (ICC2) for each SPE measure. Our findings revealed a robust experimental effect of SPE across datasets. However, when it came to individual differences, SPE measures derived from Reaction Time (RT) and Efficiency exhibited relatively higher, compared to other SPE measures, but still unsatisfied split-half reliability (approximately 0.6). Similarly, for the reliability across multiple time points, as assessed by ICC2, RT and Efficiency demonstrated low

levels of test-retest reliability (close to 0.5). These findings uncovered the presence of a reliability paradox in the context of SPMT-based SPE assessment. We discussed the implications of our findings for future studies.

Keywords: Self-Prioritization Effect (SPE), Self-Perceptual Matching Task (SPMT), 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, documented in the 1950s (Cherry, 1953), 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 objects, 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 Perceptual Matching Task (SPMT), 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 SPMT 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 SPMT as a measurement of SPE, however, has not been examined. Here, the reliability of a cognitive task refers to its consistency and dependability 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 the 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, SPMT was also employed by researchers as a measure of individual differences in SPE. For example, a recent study examined the individual differences of SPE and how these individual differences are correlated to brain network (Zhang et al., 2023). Likewise, in clinical investigation, the SPMT has been incorporated to assess deviations in self-processing among specific populations, including individuals affected by autism or depression (Hobbs et al., 2023; Liu et al., 2022). This trend calls for assessing the reliability of SPMT as a measurement of SPE.

Further, the variability in quantifying SPE using SPMT calls for a comprehensive examination of the reliability of different SPE measures. As simple as the SPMT, there are multiple approaches to quantify the SPE, encompassing various indicators and baselines. In a typical SPMT 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) 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 SPMT 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 SPMT suitable for assessing individual differences in SPE? Finally, is there a reliability paradox in the assessment of SPE using SPMT? Addressing these questions is crucial for SPMT-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 SPMT. This was achieved by re-analysing 17 independent datasets (N = 805) from 9 papers and 2 unpublished projects that employed the SPMT. In order to comprehensively assess the SPE measures derived from SPMT, we created a "multiverse" of possible indicators (RT, ACC, d-prime, d', 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) for assessing the consistency of task performance over time. The findings of our study provided valuable insights into the reliability of SPMT and its indicators, having the potential to facilitate the future utilization of SPMT 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 SPMT, as described in Experiment 1 by Sui et al. (2012). The original SPMT 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 SPMT 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 perceptual matching task. Each trial started with a fixation cross displayed in the centre of the screen for 500 ms, followed by a shape-label pairing and fixation cross for 100 ms. the screen then went blank for 1500 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.

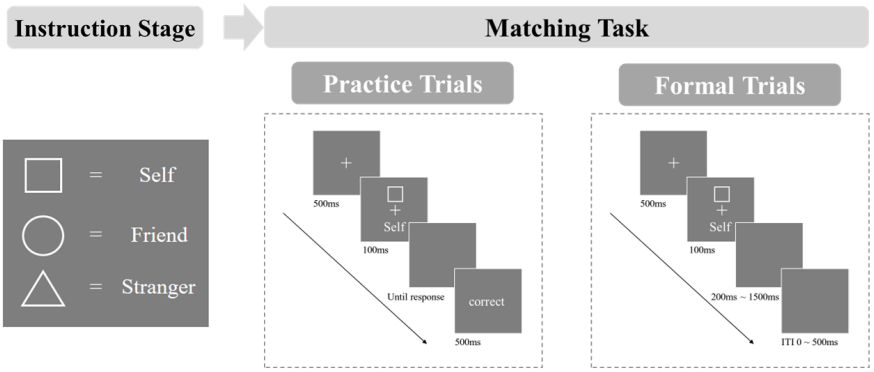

Fig. 1 Procedure of the original SPMT in Experiment 1 (Sui et al., 2012). *Note*: The relation between shape-label pairs was counter-balanced between participants.

2.3 Datasets Acquisition

Initially, two datasets that employed the SPMT 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 SPMT (pre-registration available at OSF (https://osf.io/euqmf)). During this process, we identified an additional 13 papers with datasets potentially suitable for our present study. The selection of these papers was based on specific criteria:

1) The paper must primarily utilize the SPMT 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 declared to be obtainable upon request, enabling us to estimate at least one reliability index.

Among the 13 papers included, 7 papers made their trial-level data publicly available (Constable & Knoblich, 2020; Constable et al., 2021; Golubickis & Macrae, 2021; Navon & Makovski, 2021; Qian et al., 2020; Schäfer & Frings, 2019; Svensson et al., 2022). For the remaining 6 papers, we reached out to the authors and requested access to their trial-level data. Out of those 6 requests, 3 papers provided us with trial-level data (Kolvoort et al., 2020; Woźniak et al., 2018; Xu et al., 2021). However, in one

article, the author did not provide an explanation of the shape and label in the original data (Kolvoort et al., 2020). As a result, we were unable to analyze the raw data in this context. Two papers provided us only with descriptive results (Cheng & Tseng, 2019; Martínez-Pérez et al., 2020), which unfortunately could not be used for calculating reliability. Additionally, one paper referred to data being shared on the Open Science Framework (OSF) platform https://osf.io/pcv3u/) (Bukowski et al., 2021), but we found that the repository was empty, making it ineligible for the current analysis.

In total, our analysis comprised raw data from 9 papers and 2 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 original design of SPMT without incorporating any stimuli that could potentially trigger a familiarity effect. 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 SPMT. In the end, we were able to incorporate 17 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 SPMT, 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 Datasets 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	Emotion **Control**, Neutral, Happy, Sad	Session **1-6**	33	60	√	√	√	√	√	√	√	√
Constable and Knoblich (2020)	1	Matching	Identity	Switch Identity Partner, Stranger	Phase **1-2**	92	40	√	√	√	√	√	√		√
Constable et al. (2021)	2	Matching	Identity Self; Stranger	– –	– –	56	24	√	√	√	√	√	√		√
Qian et al. (2020)	1	Matching	Identity Self; Stranger; Celebrity	Cue With, **Without**	– –	25	24	√	√	√	√	√	√		√
	2	Matching	Identity Self; Celebrity	Cue With, **Without**	– –	32	50	√	√	√	√	√	√		√
Schäfer and Frings (2019)	1	Matching	Identity Self, Mother, Acquaintance	– –	– –	35	24	√	√	√	√	√	√		√
Golubickis and Macrae (2021)	1	Matching	Identity	Presentation **Mixed**; Blocked	– –	30	30	√	√	√	√	√	√		√
Navon and Makovski (2021)	1	Matching	Identity	– –	– –	13	60	√	√	√	√	√	√		√
	3	Matching	Identity Self; Father; Stranger	– –	– –	28	60	√	√	√	√	√	√		√
	4	Matching	Identity	– –	– –	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	Tasks Modified; **Unmodified**	– –	105	60	√	√	√	√	√	√		√
Woźniak et al. (2018)	1	Matching	Identity	Facial Gender Male; Female	– –	18	56	√	√	√	√	√	√		√
	2	Matching	Identity	Facial Gender Male; Female	– –	18	60	√	√	√	√	√	√		√
Liu et al. (2023)	1	Matching	Identity Self; Stranger	– –	– –	298	16	√	√	√	√	√	√		√

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

2.4 Analysis

Analysis plans for this study were preregistered on OSF (https://osf.io/pcv3u/). 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) was 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.

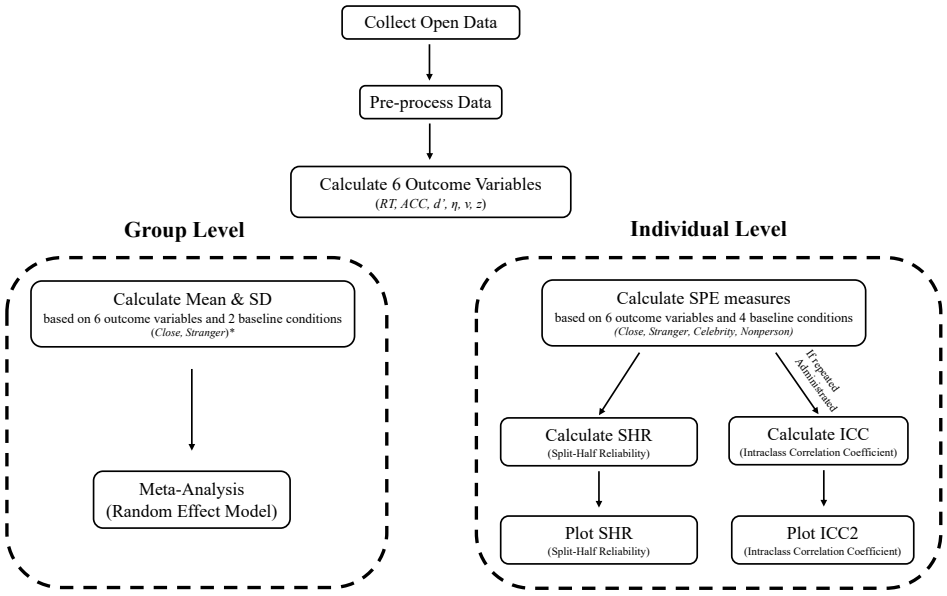

Fig. 2 Roadmap of the Current Study. *Note*: Only one paper have Celebrity and Nonpersons baseline, thus no included in the meta-analysis

2.4.1 Data Pre-processing

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

1. Participant Exclusion Criteria

   (i) Participants who had wrong trial numbers because of procedure errors is excluded from the analysis,

   (ii) participants with an overall accuracy < 0.5 is excluded from the analysis,

   (iii) participants with any of the conditions with zero accuracy is 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. 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

Indicators	Indicators Calculation	SPE Measures Calculation	Example Literature
Reaction Time (RT)	Total Reaction Time /Total Responses	$RT_{\text{self-matching}} - RT_{\text{other-matching}}$	Humphreys and Sui (2015), Sui et al. (2012), and Sui and Humphreys (2017)
Accuracy (ACC)	# of Correct Responses/Total Responses	$ACC_{\text{self-matching}} - ACC_{\text{other-matching}}$	Constable, Elekes, et al. (2019), Enock et al. (2018), and Sui et al. (2012)
d-prime (d')	$\mathcal{Z}$ (Hits) $-$ $\mathcal{Z}$ (False Alarms)	$d'_{\text{self-matching}} - d'_{\text{other-matching}}$	Hu et al. (2020) and Sui et al. (2012)
Efficiency (η)	RT/ACC	$\eta_{\text{self-matching}} - \eta_{\text{other-matching}}$	Humphreys and Sui (2015) and Stoeber and Eysenck (2008)
Drift Rate (v)	Decomposed from RT and choice based on standard DDM	$v_{\text{self-matching}} - v_{\text{other-matching}}$	Golubickis et al. (2020) and Golubickis et al. (2017)
Starting Point (z)		$z_{\text{self-matching}} - z_{\text{other-matching}}$	Macrae et al. (2017) and Reuther and Chakravarthi (2017)

Note: $\mathcal{Z}(\cdot)$ denotes the calculation of Z-score. In this context,"hit" refers to the ACC in matching trials, while "false alarm" refers to the error rate (1- ACC) in mismatch trials; the condition "Other" vary across contrast, we calculated the SPE for each "Other" condition.These could be the differences for "Self vs Close other", "Self vs Stranger", "Self vs Celebrity" or "Self vs Non-person".

2.4.3 Estimating the Robustness of SPE

The robustness of experimental effects (group-level effect) of SPE in SPMT 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 rigourness, 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.

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 four splitting approaches (Pronk et al., 2022). When using the Monte Carlo 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, odd-even split, and permutated split were similar to the Monte Carlo approach except each of these approaches 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 synthesized these reliability coefficients via a meta-analytical approach. We weighed the reliability coefficients based on the trial numbers of

each study since the number of trials typically significantly influences the reliability of cognitive experiments (Kucina et al., 2023) (see also Supplementary Fig. S7 for our exploratory analysis). The weighted-average reliabilities were calculated use the "aggregate.escalc" function in the "metafor" Package (Viechtbauer, 2010). We reported the synthesized split-half reliability and its 95% confidence interval in the results section. Although there is no strict criterion for defining the level of split-half reliability for psychological and educational measures, a widely accepted guideline for split-half reliability coefficient is that a value of 0.5 is "poor", a value of 0.70 is "acceptable", and a value greater than 0.8 means 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 SPMT 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 (ICC2) within the ICC family (Chen et al., 2018; 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:

$$\text{ICC2} = \frac{MSBS - MSE}{MSBS + (k-1)MSE + \left(\frac{k}{n}\right)(MSBM - MSE)}, \tag{1}$$

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 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 favourable 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, considering the significant impact of the number of trials on reliability (Kucina et al., 2023), during the formal analysis, we assigned different weights to each study based on the number of trials. Subsequently, we calculated a weighted average of the split-half reliabilities. 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, Monte Carlo split-half reliability, and effect size (Hedges' g) (refer to Supplementary Fig. S7-9). Finally, 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. In this final report, we improved the writing and adjusted the introduction and abstract accordingly.

4 Results

Of 17 independent datasets, 14 of them contain data for "Close other", 13 of them contain data for "Stranger", 1 of them has the data for "Celebrity", and 1 of them has the data for "Nonperson". Since there is only one paper for "Celebrity" and "Nonperson", their results were less robust and 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 SPMT. 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 SPMT

Baseline	Indicators	Hedges' g[95%CI]	# of Studies	Q	p	I^2
Close other	RT	0.47 [0.30, 0.63]	14	68.67	$< .001$	84.94%
	ACC	0.73 [0.42, 1.03]	14	144.57	$< .001$	92.87%
	d'	0.44 [0.28, 0.59]	14	81.96	$< .001$	83.02%
	η	0.88 [0.50, 1.25]	14	128.47	$< .001$	94.67%
	v	0.54 [0.32, 0.76]	14	142.79	$< .001$	91.16%
	z	0.15[−0.03, 0.33]	14	122.30	0.11	88.95%
Stranger	RT	0.59 [0.40, 0.78]	13	55.30	$< .001$	83.20%
	ACC	0.78 [0.48, 1.08]	13	77.78	$< .001$	88.60%
	d'	0.35 [0.21, 0.50]	13	47.81	$< .001$	75.38%
	η	0.92 [0.56, 1.29]	13	98.79	$< .001$	93.30%
	v	0.44 [0.28, 0.59]	13	50.98	$< .001$	79.33%
	z	0.08[−0.09, 0.24]	13	70.48	0.37	84.44%

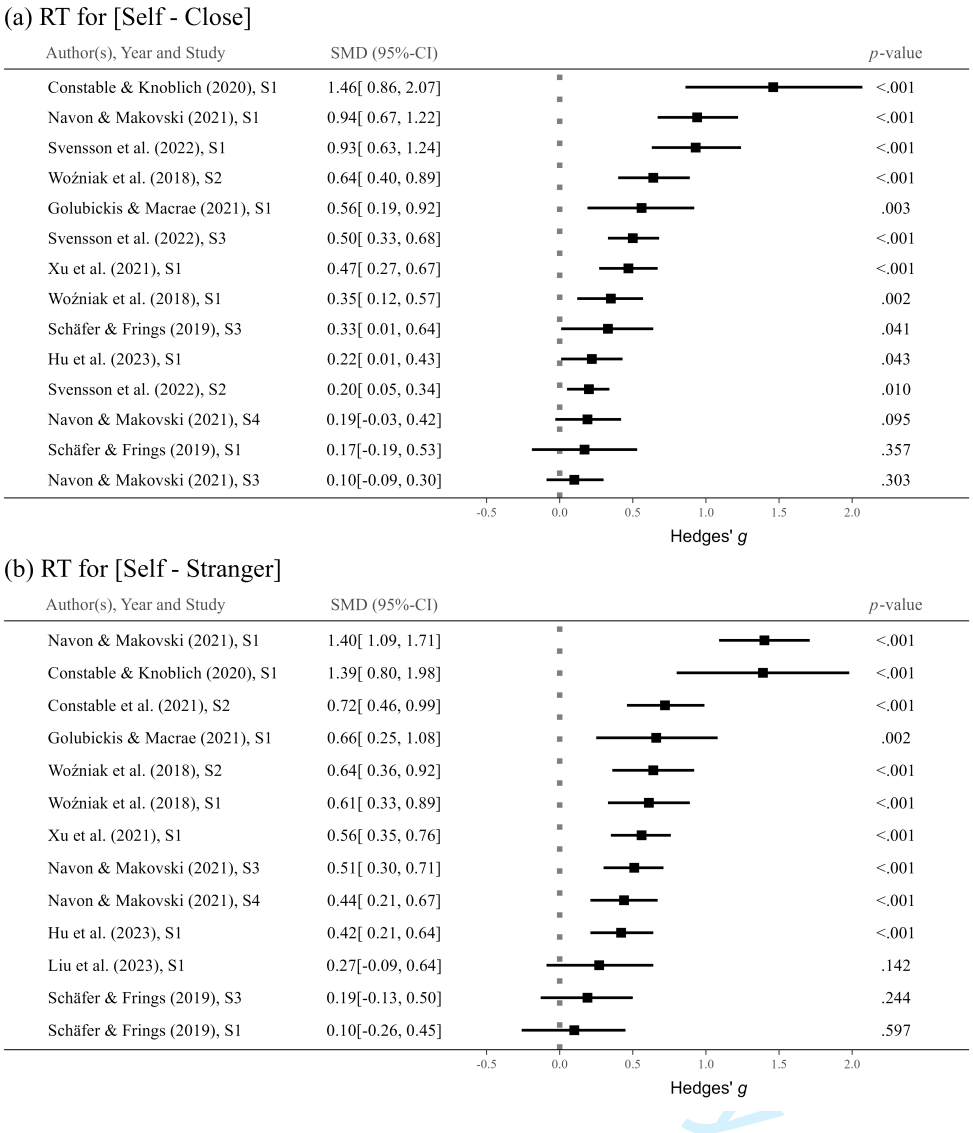

Fig. 3 Forest Plots for Group-level Self-Prioritization Effect (SPE) as Quantified by RT. (a) When "Close other" as the baseline condition for SPE, i.e., the "Self - Close other" contrast; (b) When "Stranger" as the baseline condition for SPE, i.e., the "Self -Stranger" contrast.

4.2 Split-half Reliability

We used four different approaches to split the data when calculating split-half reliability: the first-second, odd-even, permuted, and Monte Carlo methods. Also, we used the weighted average split-half reliability as the overall reliability across studies. Here we only presented the results from the Monte Carlo 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 three split-half methods can be found in the supplementary materials (see Supplementary Fig. S4).

We found that, among all SPE measures, the four with highest split-half reliabilities were as follows: Reaction Time (RT) with "Stranger" as baseline ($r = .65, SE = .02, p < .001, 95\%\text{CI}$ [.61, .70]); Efficiency (η) with "Stranger" as baseline ($r = .64, SE = .03, 95\%\text{CI}$ [.59, .69]); RT with "Close other" as baseline ($r = .58, SE = .02, 95\%\text{CI}$ [.54, .63]); η with "Close other" as baseline ($r = .57, SE = .02, 95\%$ CI [.52, . 62]). These SPE measures achieved a split-half reliability of around 0.6 or higher, which is considered acceptable. 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 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 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 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 SPMT, the result showed a similar pattern here (see Supplementary Fig. S5).

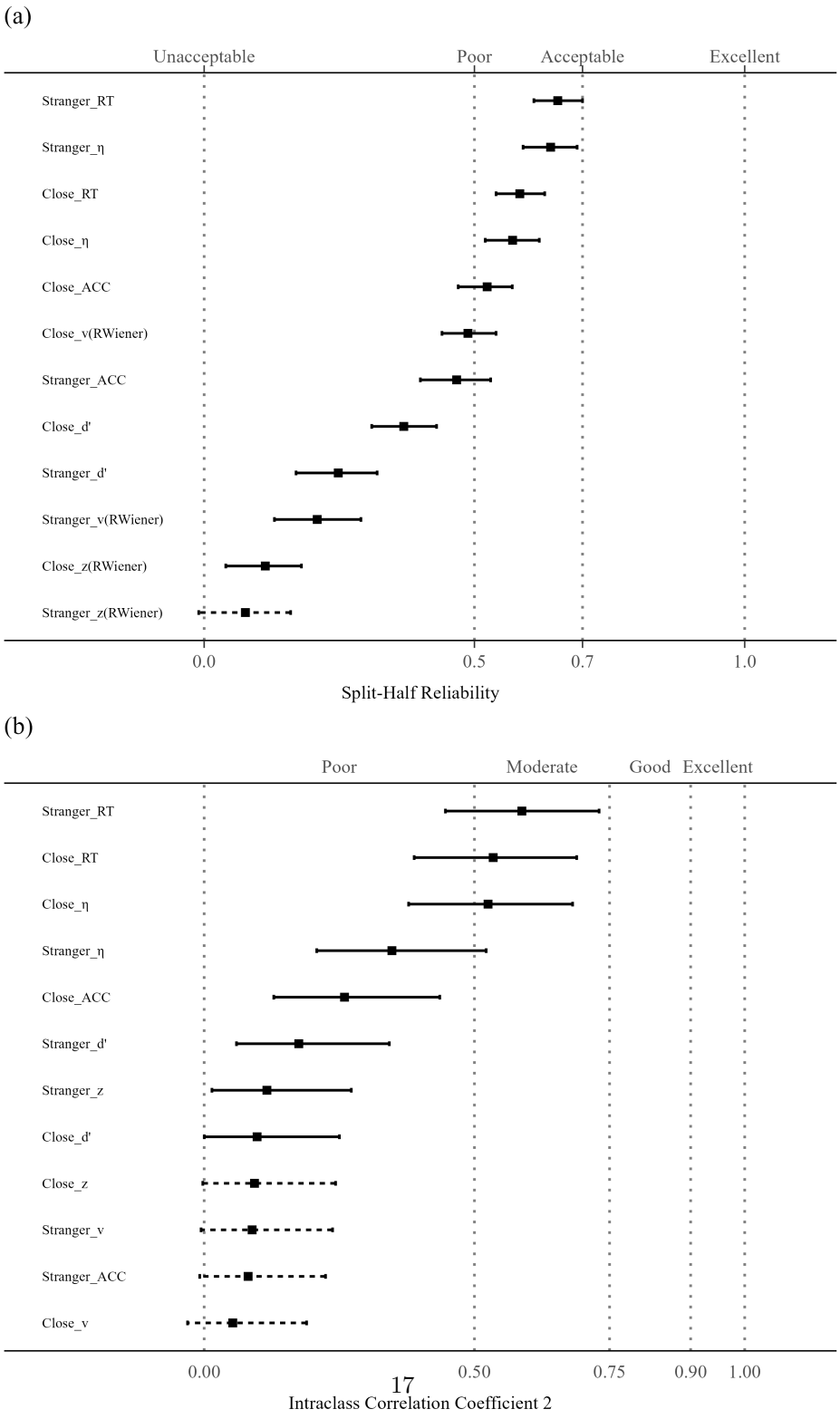

Fig. 4 Reliability for Different SPE Measures. (a) The Weighted Average Split-Half Reliability (Monte-Carlo); (b) Intraclass Correlation Coefficient (ICC2). *Note*: The vertical axis represents 12 different SPE measures, combining six indicators (RT, ACC, d', η, v, z) and two baseline conditions ("Close 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 non-significant value. Due to the fact that there is only one paper for "Celebrity" and one for "Nonperson", their results is presented in the supplementary materials.

5 Discussion

In this pre-registered study, we examined the reliability of various measures from the Self Perceptual Matching Task (SPMT) in assessing the self-prioritization effect (SPE) using a multiverse approach. Our analyses revealed that except for parameters z from DDM, all the other measures exhibited robust SPE. However, when it came to reliability, only two measures of SPE, Reaction Time and Efficiency, exhibited acceptable to moderate reliability, among all indicators that have been reported in the literature. Our results suggested that the current implementation of SPMT was not well-suited for assessing individual differences. Taken together, our study revealed a "reliability paradox" of SPE as measured by SPMT. These findings provided important methodological insights for future studies of SPE.

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 group level, effect sizes of SPE as measured by Reaction Time and Efficiency were moderate to large effect; on 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. Moreover, 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. Therefore, 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" in SPMT. 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 SPMT warrant thorough investigation. However, one of the most plausible explanations is that the SPMT, 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). Consequently, this reliability paradox sheds light on the specific types of inquiries that the SPMT can proficiently address and those it cannot.

More specifically, the relatively low reliability of all the SPE measures calls for attention when researchers are interested in measuring individual differences, such as in clinical settings (e.g., Karvelis et al., 2023), or searching an association with data from questionnaires (Hedge et al., 2018)). As the SPMT was designed to achieve robust group-level SPE rather than to measure individual differences, researchers need to

re-design the task if they are interested in assessing individual differences. Recently, researchers have proposed several ways to enhance the reliability of cognitive tasks, such as gamification (Friehs et al., 2020), using 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 concerns about applying the standard drift-diffusion model to data from SPMT directly. Previous studies also found that the standard drift-diffusion model did not fit the data from matching task (Groulx et al., 2020). 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 DDM blindly.

5.1 Implications of the Current Study

Our findings can offer an initial guide for researchers considering the use of SPMT. 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. 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, our results reported here assessed the robustness and reliability of SPE with the original experimental design of Sui et al. (2012), which means the robustness and reliability of different variants of SPMT still need further investigation. For a more systematic meta-analysis of SPE measured by SPMT, please see our ongoing project (https://osf.io/euqmf). 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 an empirical assessment of the reliability of the self-perceptual matching task (SPMT). We found a robust self-prioritization effect for Reaction Time and Efficiency. Meanwhile, the reliability of the most robust SPE measure fell short of being satisfactory. The results of the current study may serve as a benchmark for the improvement of future studies.

Acknowledgments

The data collection from Hu et al. (2023) was supported by the National Science Foundation (China, Grant No. 31371017) to JS.

Author Contributions

HCP contributed to the conception and supervision of the study. HCP contributed to data collection of Hu et al. (2023). JS contributed to funding acquisition. LZ, ZYR and HMZ wrote the simulation code for pre-registration. HMZ collected the datasets from published papers and performed data pre-processing, analysis and visualisation of the results. LZ, HMZ and HCP contributed to discussing the results and the drafting of the final manuscript. HCP, JS, LZ and HMZ critically revised the manuscript.

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

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

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. *Am J Ment Defic*, *86*(2), 127–137. https://psycnet.apa.org/record/1982-00095-001

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

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

Fisher, R. A. (1992). Statistical methods for research workers. *Springer New York*. https://doi.org/10.1007/978-1-4612-4380-9_6

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

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. 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, 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., Bathke, A., & Blechert, J. (2022). Reliability of reaction time tasks: How should it be computed? 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/S0926-6410(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/s41467-023-37777-2

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

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 ego-depletion paradigm. *Psychological Science*, *31*(5), 531–547. https://doi.org/10.1177/0956797620904990

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, Sui, J., & Hildebrandt, A. (2023). To see or not to see: The parallel processing of self-relevance and facial expressions. *Manuscript submitted for publication.*

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, double-blind, and multichannel hd-tdcs study. *Frontiers in Neuroscience*, *14*, 683. https://doi.org/10.3389/fnins.2020.00683

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

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

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

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

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *J Pers Soc Psychol*, *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

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., 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

Svensson, S. L., Golubickis, M., Maclean, H., Falbén, 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

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 providing 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 modeling 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

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/s12144-021-02114-7

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. 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 Perceptual 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 Sciences, 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.

Contents

List of Figures

1 Supplementary Methods

1.1 Methodological details of 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 SPMT with large deviations), Experiment B (another modified SPMT 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 ($\text{Mean}_{\text{age}} = 21.06$, $\text{SD}_{\text{age}} = 3.24$), 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) $\times$ 3 (Idnetity: Self, Friend, Stranger) $\times$ 4 (Emotion: Control, Neutral, Happy, Sad) $\times$ 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 $\times$ 768 resolution with 100 Hz refresh rate.

The experiment has two phases (see Fig. S1). Following Sui et al. (2012), the first phase comprised an instruction stage in which participants were required to associate geometric shapes with labels. The shapes were not presented at this stage. The instruction stage lasted for approximately 60 seconds and shape-target associations

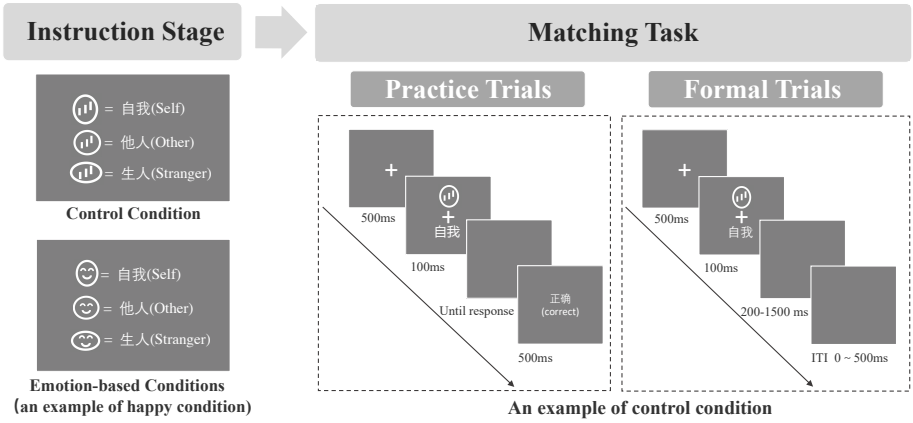

Fig. S1 Procedure of the SPMT in Experiment B (Hu et al., 2023). *Note*: The labels and feedback appeared in Chinese in the experiment. In the associative learning task, the matched associations of shapes and labels was counterbalanced between participants. Timely feedback was not provided in formal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

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 centre of the screen for 500 ms. Then, a shape–label pairing as well as the fixation cross was presented for 100ms, respectively. The next frame showed a blank screen for 1500 ms, or until a response was made. 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.

1.2 Parameter Recovery Results 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. S2 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.

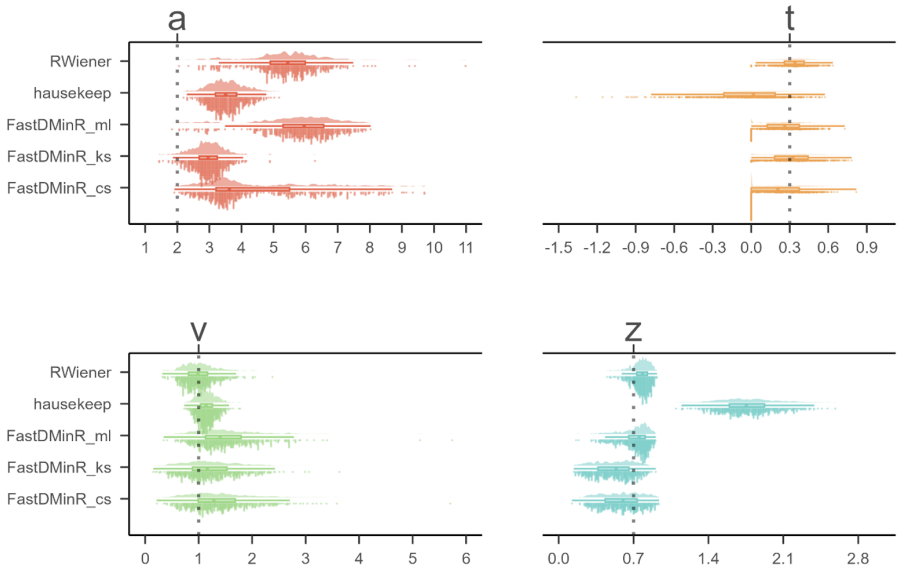

Fig. S2 DDM Packages Comparison. *Note*: The parameters of interest in the Drift-Diffusion Model (DDM) are represented as follows: "a" denotes the threshold parameter, "t" represents the non-decision time, "v" indicates the drift rate, and "z" corresponds to the starting point. The y-axis of the graph 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 were presented in Fig. S3.

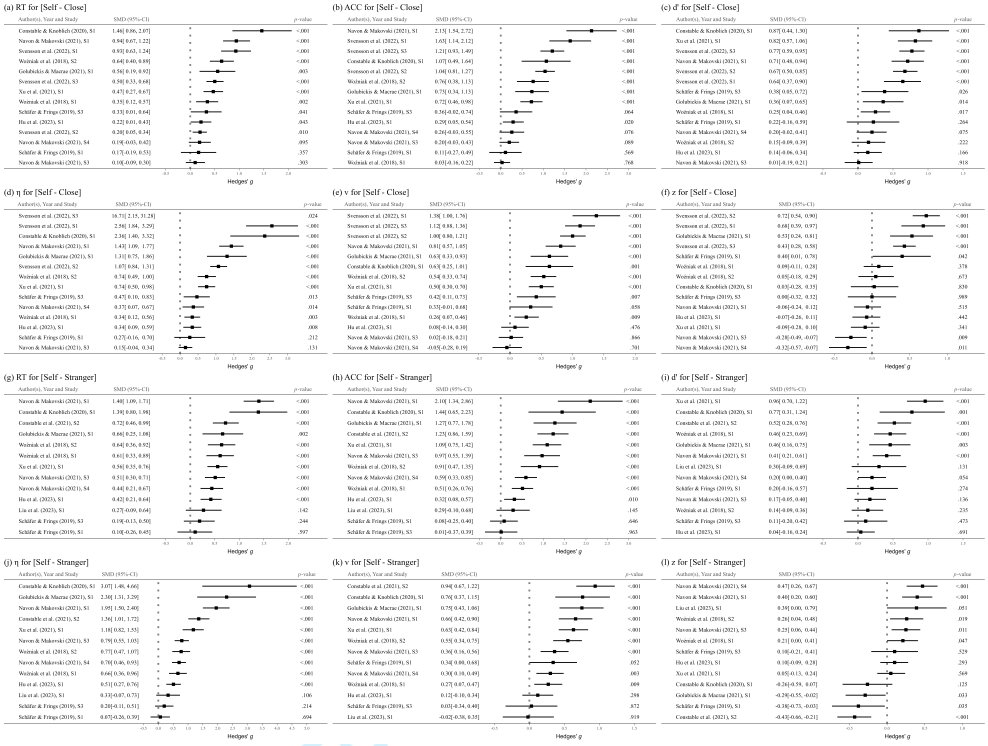

Fig. S3 Forest Plot for SPE Measures. *Note*: Fig (a)-(f) represent the forest plots corresponding to RT, ACC, d', η, v, and z under the condition where Target is Close. Fig (g)-(l) represent the forest plots corresponding to d', η, v, and z under the condition where Target is Stranger.

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 SPMT

Baseline	Indicators	Hedges' g[95%CI]	t	df	p
Celebrity	RT	1.76 [1.11, 2.41]	5.28	24	$< .001$
	ACC	2.08 [1.39, 2.77]	5.93	24	$< .001$
	d'	1.41 [.79, 2.03]	4.45	24	$< .001$
	η	2.70 [1.93, 3.46]	6.90	24	$< .001$
	v	1.45 [.83, 2.08]	4.57	24	$< .001$
	z	.05 [−.50, .61]	0.19	24	.85
NonPerson	RT	.13 [−.36, .62]	-.51	31	.61
	ACC	.02 [−.47, .51]	.07	31	.95
	d'	.17 [−.32, .66]	.68	31	.50
	η	.09 [−.40, .58]	-.36	31	.72
	v	.33 [−.16, .83]	1.32	31	.19
	z	−.45 [−.95, .04]	-1.79	31	.07

2.2 Split-Half Reliability Using Four Splitting Approaches

In this section, we presented the Split-Half Reliability (SHR) results for the SPE measures using four split-half methods: Monte Carlo, 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. S4 for the visual representation of the results.

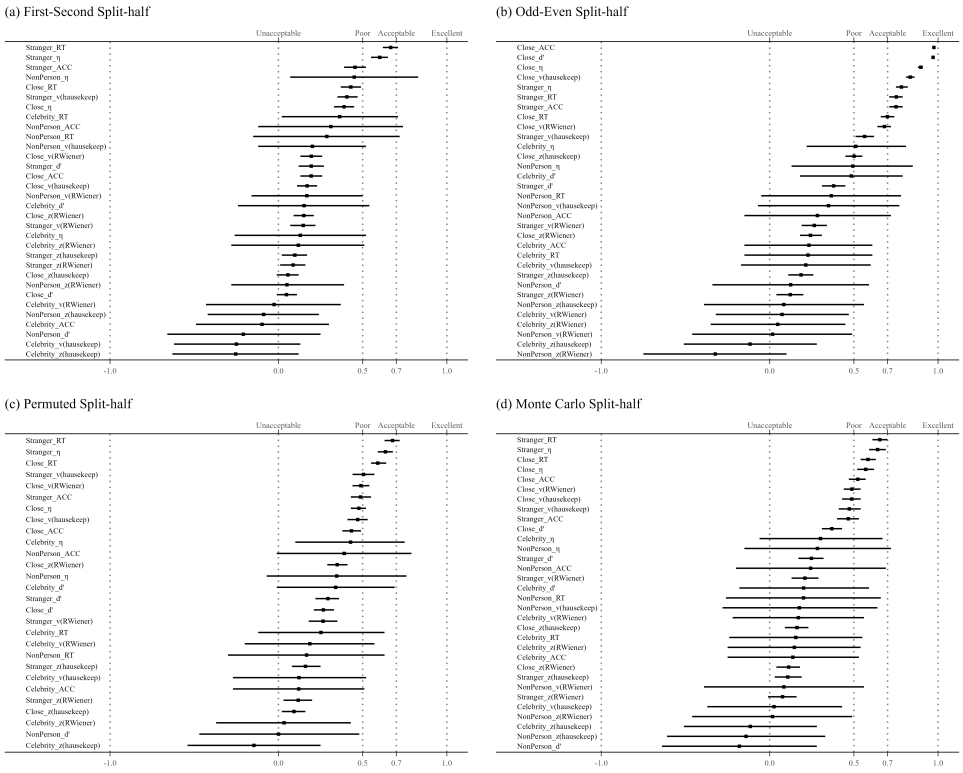

Fig. S4 Results of SHR Using Four Split-half Methods. (a) Results of SHR using First-Second Split-half Methods; (b) Results of SHR using Odd-Even Split-half Methods; (c) Results of SHR using Permuted Split-half Methods; (d) Results of SHR using Monte Carlo Split-half Methods. *Note*: The vertical axis 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 permuted.

It's evident that the pattern of the results from the permuted split-half methods and first-second split-half methods closely resembled the Monte Carlo 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 three 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 SPMT. 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. S5, 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 SPMT used in this dataset deviated quite strongly from the original SPMT paradigm. Due to these significant differences, ICC2 obtained from this dataset may reflect variations introduced by the modified SPMT rather than directly comparable results to the original paradigm.

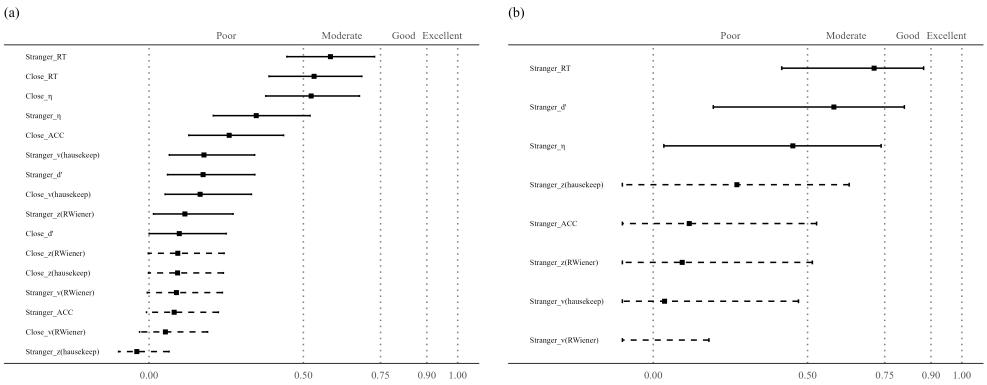

Fig. S5 ICCs for SPE Measures Using Hu et al. (2023) and Another Dataset. (a) ICC2 for SPE measures using Hu et al. (2023); (b) ICC2 for SPE measures using an additional dataset. *Note*: The vertical axis of the graph illustrates eight distinct indicators, which includes two additional indices from the 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. S6, 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.

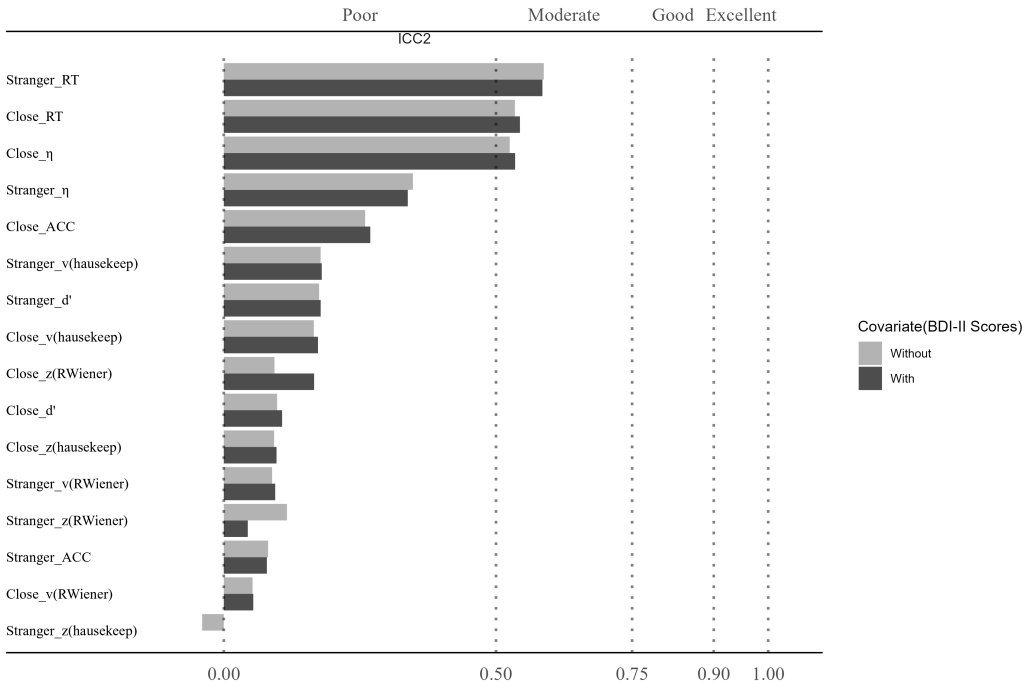

Fig. S6 ICC2 for SPE Measures Using Hu et al. (2023) with Covariant (BDI-II Scores). *Note*: The vertical axis of the graph illustrates eight distinct indicators, which includes two additional indices from the DDM, implemented using the "hausekeep" package. The bar on the graph represent the value of ICC2.

2.4 Exploratory Analyses

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: Monte Carlo split-half reliability and effect size (Hedges' g). We also examine the relationship between Monte Carlo split-half reliability and effect size (Hedges' g).

We found significant correlations between trial numbers and Monte Carlo split-half reliability for some indicators, such as Reaction Time and Efficiency (see Fig. S7) . However, for indicators like d' and v, the correlation with trial numbers was relatively weak. Moreover, we could observe that the SPMT paradigm requires approximately 80 trials to achieve a Monte Carlo split-half reliability of 0.8 for the SPE measure of RT under the "Stranger' condition and around 120 trials under the "Close' condition. Furthermore, achieving a Monte Carlo SHR of 0.8 of the parameter v may require more than 120 trials. On the other hand, attaining high Monte Carlo SHR values for the remaining three indicators, particularly for the z parameter, remains challenging even with 150 or more trials.

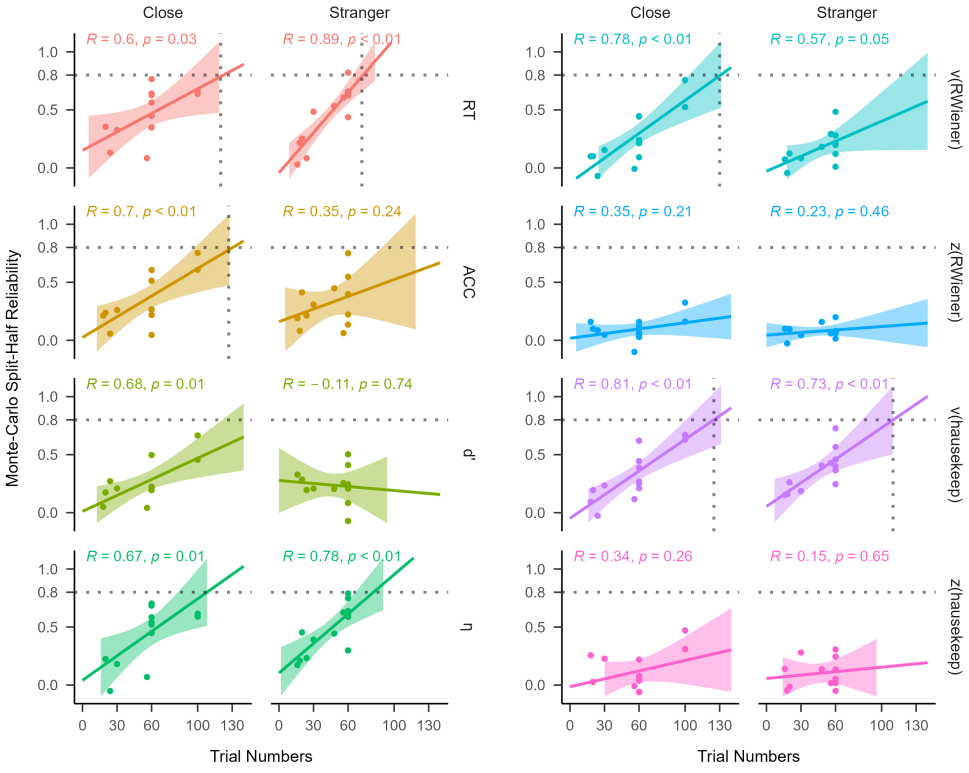

Fig. S7 Regression Analysis Between Monte Carlo SHR and Trial Numbers Using Different SPE Measures. *Note*: The vertical axis represents Monte-Carlo split-half reliability, and the horizontal axis represents the number of trials. Each facet represents one SPE measures.

We also explored the correlation between split-half reliability and effect size (Hedges' g), as shown in Fig. S8. Our exploratory analysis did not find a significant correlation among them. This result pattern was somehow consistent with the reliability paradox (Hedge et al., 2018; Logie et al., 1996), which suggested that robust experimental effects are not always associated with robust individual difference correlations.

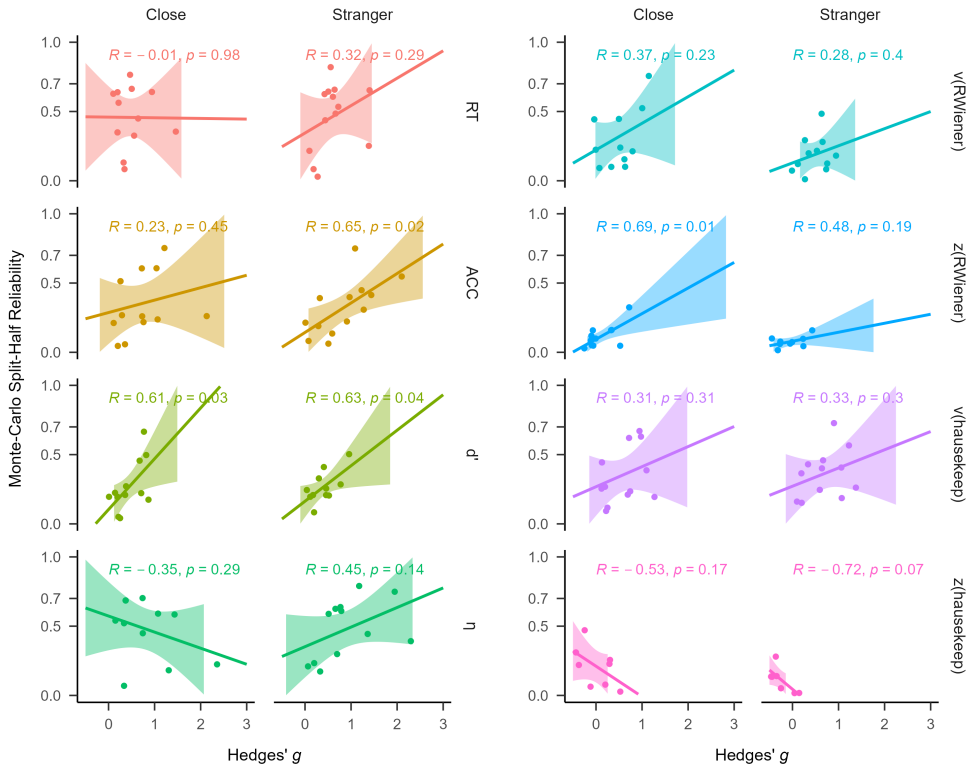

Fig. S8 Regression Analysis Between Monte Carlo SHR and Effect Size (Hedges' g) Using Different SPE Measures. *Note*: The vertical axis represents Monte-Carlo split-half reliability, and the horizontal axis represents the effect size (Hedges' g). Each facet represents one SPE measures.

Finally, we calculated the correlation coefficient between trial numbers and effect size (Hedges' g), as shown in Fig. S9. Similarly, no significant correlation was found.

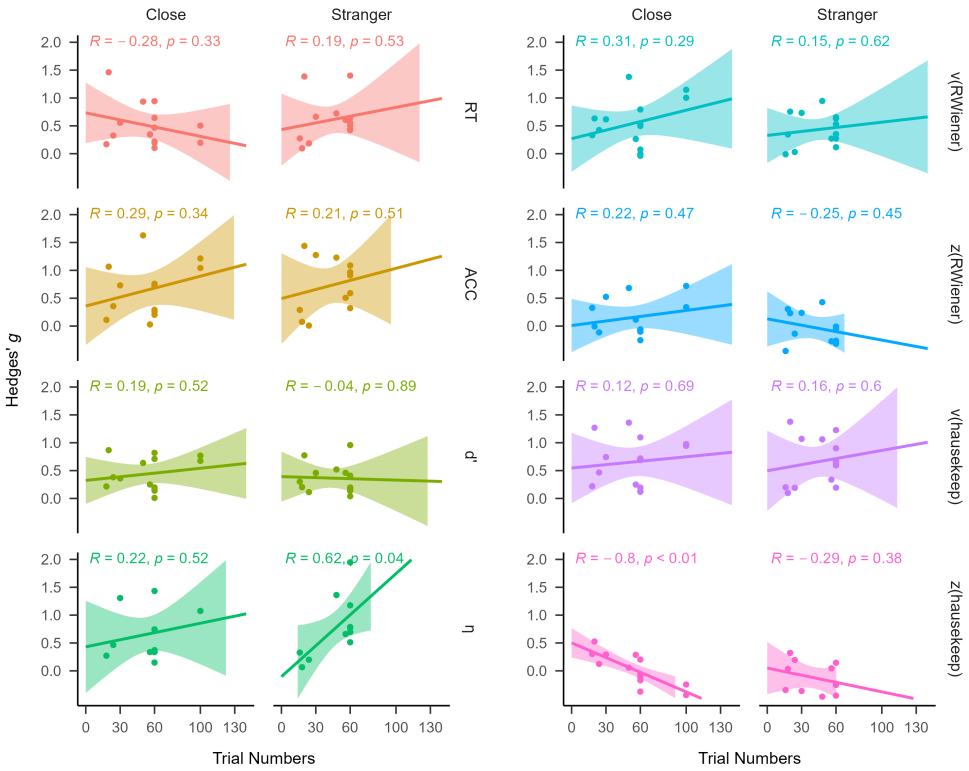

Fig. S9 Regression Analysis Between Trial Numbers and Effect Size (Hedges' g) Using Different SPE Measures. *Note*: The vertical axis represents the effect size (Hedges' g), and the horizontal axis represents trial numbers. Each facet represents one SPE measures.

It's important to emphasize that here we only conducted a simple regression analysis of these variables. This analysis was not part of the pre-registered plan, and our primary aim was not to provide a well-validated improvement for the SPMT. Nevertheless, taking into account the noteworthy correlation observed between the number of trials and Monte Carlo split-half reliability, our results indicated that when employing the SPMT 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

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