**Reliability Assessment of Self-Prioritization Effect as Measured by the Perceptual Matching Task**

Zheng Liu1,2#, Mengzhen Hu1#, Yuanrui Zheng1, Jie Sui3, Hu Chuan-Peng 1\*

1 School of Psychology, Nanjing Normal University, Nanjing, China

2 School of Humanities and Social Sciences, The Chinese University of Hongkong (Shenzhen), Shenzhen, China

3 School of Psychology, University of Aberdeen, Old Aberdeen, Scotland

# These authors are equally contributed to this study

\* Corresponding authors: Hu Chuan-Peng ([hu.chuan-peng@nnu.edu.cn；hcp4715@hotmail.com](mailto:hu.chuan-peng@nnu.edu.cn；hcp4715@hotmail.com))

***Abstract:***

The self-prioritization effect (SPE) refers to the phenomenon where individuals demonstrate faster and more accurate responses to self-relevant stimuli in cognitive tasks. The self-perceptual matching task (SPMT) has emerged as a mainstream paradigm for studying the cognitive mechanisms underlying the SPE. Additionally, the SPE measured by the SPMT has been utilized as an indicator of individual differences in research, including clinical studies. As a simple button-pressing task, SPMT eliminates familiarity effects and yields two outcome variables for quantifying SPE: reaction time (RT) and accuracy (ACC). Existing literature has also reported four additional outcome variables derived from reaction times and accuracy, including sensitivity *d’* under signal-detection theory, the efficiency (*η*), drift rate (*v*) and starting point (*z*) estimated using the drift-diffusion models. Moreover, the calculation of SPE can employ different baseline conditions, including “Close other”, “Stranger”, “Celebrity”, and “Non-person”. However, the reliability of SPE measures computed using different baseline conditions and outcome variables has not been systematically examined, leaving the stability of using the SPMT for measuring individual differences unknown. To fill the gap, we conducted a pre-registered study wherein we re-analyzed 18 datasets from 9 papers and two projects (N = 857) using the weighted average split-half reliability (WASR) and intraclass correlation coefficient (ICC). The results revealed that the split-half reliabilities of RT (WASRClose = 0.58; WASRStranger = 0.60) and Efficiency (WASRClose = 0.52; WASRStranger = 0.58), when close other and stranger used as the baseline condition, are relatively high but still lower than that required in psychometrics. The estimated WASR for all other indices are approximately 0.5 or even lower. Similar results were obtained in ICCs, where the ICC2 for individual differences in the reaction time measures (ICC2Close = 0.53; ICC2Stranger = 0.58) and efficiency measures (ICC2Close = 0.52; ICC2Stranger = 0.34) are relatively high when using close other or stranger as baseline conditions, but still considerably lower than the desired level in psychometrics. For the ICC2k, which measures the stability of SPE on group level, the reaction time measures (ICC2kClose = 0.87; ICC2kStranger = 0.89) and efficiency measures (ICC2kClose = 0.86; ICC2kStranger = 0.76) are relatively high when using close other or stranger as baseline conditions, while the remaining 10 ICC2k indices are about 0.5 or even lower. These findings suggest that SPE measures based the reaction time and efficiency are reliable only at the group level, while further exploration is needed for measuring individual differences. These results provide insights into the reliability of the SPMT and its future applications.

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

# **1 Introduction**

The Self-Prioritization Effect (SPE) refers to the phenomenon whereby performance in cognitive tasks is better when stimuli are related to the self than when they are not. This effect has been widely documented and confirmed since the 1950s. In the early days of cognitive psychology, researchers found that subjects were able to recognize their own names, even when they were mixed with a noisy auditory background and not the target of the task in dichotic listening tasks (Cherry, 1953; Moray, 1959). SPE was then reported in memory research by Craik and Tulving (1975), who found that participants were able to recall more words when they were related to the self compared to when they were processed at other levels (e.g., semantic). This SPE effect in memory was also replicated by many others (Conway & Dewhurst, 1995; Rogers et al., 1977; Symons & Johnson, 1997). In the following decades, the SPE has also been found to occur with different stimuli, such as own face (Keenan et al., 2000; Kircher et al., 2000; Turk et al., 2002), own voice (Hughes & Harrison, 2013; Payne et al., 2021), own name (Constable, Rajsic, et al., 2019), and newly owned object (Strachan et al., 2020). SPE was also found across a variety of cognitive tasks, such as perceptual task (Cunningham & Turk, 2017; Desebrock et al., 2018), decision-making task (Sui & Humphreys, 2013), attentional task (Shapiro et al., 1997), and ownership task (Cunningham et al., 2008).

Although SPE is often argued to be a self-specific effect, it can be challenging to disassociate it from the familiarity effect since most studies use stimuli owned by participants or by others. Sui et al. (2012) introduced the Self-Perceptual Matching Task (SPMT), a cognitive paradigm designed to investigate the acquisition of social meaning attributed to different geometric shapes. In this task, participants are initially required to establish associations between specific geometrical shapes (e.g., triangle, square, and circle) and corresponding labels of persons (e.g., "You," "friend," and "stranger"). Subsequently, participants engage in a perceptual matching task, wherein they evaluate whether the shape-label pairs presented on the screen align with the previously learned associations. In this task, Sui et al. (2012) found that shapes associated with the self showed superior performance, with faster response times, better accuracy and higher sensitivity scores (*d’*), compared to shapes associated with friends and strangers.

The most significant contribution of the SPMT lies in its successful elimination of the familiarity effect by strategically inducing self-relatedness just prior to the commencement of the perceptual matching task. As a result of this innovation, the SPMT has become the mainstream method for investigating the mechanism underlying the SPE. For instance, researchers have explored the importance of personality traits in identity labels (Golubickis et al., 2020), the self-relevant labels that include the past, present, and future self (Golubickis et al., 2017), self-related labels with moral values (Hu et al., 2020), and the group advantage effect of in-group labels (Constable, Elekes, et al., 2019; Constable & Knoblich, 2020; Enock et al., 2018; Enock et al., 2020). Moreover, the SPMT has been applied to various fields. In neuroscience and physiology, researchers investigate which brain regions are activated during the SPE (Feng et al., 2018; Humphreys & Sui, 2015), and gender differences in SPE due to oxytocin (Feng et al., 2020). The SPMT has also been applied to child development, with studies examining developmental changes in SPE (Maire et al., 2020; Zhou et al., 2019). Cross-cultural studies have demonstrated variations in the magnitude of the SPE across different cultural contexts, particularly between individualistic and collectivist cultures (Jiang et al., 2019). Finally, in clinical research, SPMT has also been integrated to understand atypical self-processing in populations such as those with autism or depression (Gillespie‐Smith et al., 2018; Nijhof & Bird, 2019; Sui & Humphreys, 2017).

Evaluating the reliability of a behavioral paradigm is essential for researchers planning to use the paradigm to investigate different research questions, such as individual differences and underlying mechanisms (Green et al., 2016; Hedge et al., 2018; Parsons et al., 2019). While SPMT has gained widespread adoption as a prominent method for investigating the underlying mechanism of SPE, there has been microscopic examination and report of the psychometric properties of the outcomes, leaving the reliability of using the SPMT for measuring individual differences and group level consistency unknown (Zorowitz & Niv, 2023). Furthermore, in tasks as simple as the SPMT, there are multiple approaches to quantify the SPE. The SPMT yields two direct outcome variables for quantifying SPE: mean reaction time (RT) and accuracy (ACC). In addition to these, the existing literature has identified four additional outcome variables derived from RT and ACC. These include efficiency (*η*) (Humphreys & Sui, 2015; Stoeber & Eysenck, 2008), sensitivity score (*d’*) under signal detection theory (Hu et al., 2020; Sui et al., 2012), and drift rate (*v*) and starting point (*z*) estimated using the drift-diffusion models (DDM) (Golubickis et al., 2017). Moreover, the calculation of SPE can employ different baseline conditions, such as “Close other” (Navon & Makovski, 2021; Svensson et al., 2022), “Stranger” (Constable et al., 2021; Orellana-Corrales et al.), “Celebrity” (Qian et al., 2020), and “Non-person” (Schäfer & Frings, 2019). Given the increasing use of SPMT to assess individual differences in fields such as psychiatry (Liu et al., 2022) and social psychology (Enock et al., 2018), it is crucial to ensure a high degree of measurement consistency to accurately assess human perceptual capacities. As a result, two pivotal questions remain unresolved: (1) Can the six outcome variables (RT, ACC, *d’*, *η*, *v*, *z*) reliably measure the SPE under various baseline conditions? Additionally, which of these outcome variables demonstrate the highest level of reliability? (2) Is the SPMT suitable for assessing individual differences and group-level variations in the manifestation of the SPE? Addressing these questions is crucial for establishing the reliability of SPMT measurements, allowing for accurate assessment of the SPE and its implications in various domains.

To address the existing research gap, the present study investigated the reliability of SPE measures computed using different outcome variables under various baseline conditions in the SPMT. This was achieved by re-analyzed data of 18 independent studies from 9 papers and 2 projects (N = 857) that employees SPMT. In order to comprehensively assess the SPE measures derived from SPMT, we examined six outcome variables ((RT, ACC, *d’*, *η*, *v*, *z*) under four baseline conditions (“Close other”, “Stranger”, “Celebrity”, and “Non-person”), as mentioned earlier, that is supposed to capture the disparity between self-related and other-related stimuli of the matching trials. Given the methods available for evaluating the reliability of cognitive tasks, we employed both the Split-Half Reliability (SHR) and Intraclass Correlation Coefficient (ICC) to determine the reliability of each SPE measures. The findings of our study aim to provide valuable insights into the reliability and consistency of SPMT and its outcome variables, having the potential to facilitate the future utilization of SPMT in research, clinical settings, and personal performance monitoring.

# **2 Methods**

## 2.1 Ethics information

Since this research involves a secondary analysis of pre-existing data obtained from publicly available datasets or archived data from author’s group, which have used SPMT in recent years, informed consent and confidentiality are not applicable.

## 2.2 Experimental Design

We first provided a brief overview of the original experimental design of SPMT, as described in the Experiment 1 by Sui et al. (2012). The original SPMT used a 2 by 3 within-subject design. The first independent variable, labeled “Matching,” consisted of two levels: “Matching” and “Nonmatching,” indicating whether the shape and label were congruent. The second independent variable, labeled “Identity”, comprised three levels: “Self”, “Friend”, and “Stranger”, representing the corresponding identity associated with the shape.

The original SPMT consisted of two stages (see Fig. 1). In the first stage (learning stage), participants completed a learning task in which they associated three geometric shapes (circle, triangle and square) with three labels (self, friend, and stranger) for approximately 60 seconds. The shape-label associations were counter-balanced across participants. In the second phase (formal experimental phase), participants completed a perceptual 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 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 allotted 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/nonmatching) 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 the Experiment 1 (Sui et al., 2012).

## 2.3Datasets Acquisition

At the beginning, we had access to two datasets that utilized the SPMT. One dataset belonged to our laboratory (Hu et al., 2023), and the other came from our collaborators (Liu et al., 2023). As part of an ongoing meta-analysis on SPE using SPMT (see pre-registration on https://osf.io/euqmf), we identified other 13 papers that have potential usable datasets for this study. These papers were selected based on the selection criteria:

1. The paper must primarily utilize the SPMT as their method.
2. The experimental design should align with the original SPMT by Sui et al. (2012).
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.

During our investigation, we discovered that seven papers made their raw data publicly available (Constable et al., 2021; Constable & Knoblich, 2020; 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 six requests, three papers provided us with usable trial level data (Kolvoort et al., 2020; Woźniak et al., 2018; Xu et al., 2021); two papers provided us only with descriptive results, which unfortunately could not be used for calculating reliability (Cheng & Tseng, 2019; Martínez-Pérez et al., 2020); one article 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 inclusion in the current analysis.

In total, our analysis comprised raw data from 9 papers, along with 2 project datasets contributed by our laboratory and collaborator. It is important to highlight that the research culture we encountered discourages direct replications (Makel et al., 2012). As a result, all the datasets included in our analysis underwent some degree of modification to the original design, including the incorporation of additional independent variables or the use of different experimental materials (refer to Table 1 for specification). For our analysis, we exclusively utilized the data from the baseline condition without incorporating these additional independent variables. Thus, we were able to incorporate 18 independent studies from these papers and projects. Nonetheless, not all studies incorporated repeated measures. If a publicly available datasets did not include repeated measurements using SPMT within a specified time interval, we excluded it from calculating the Intraclass Correlation Coefficient (ICC) and only considered split-half reliability. The details of the included studies and conditions in the datasets are described in Table 1.

Table 1. Dataset Information

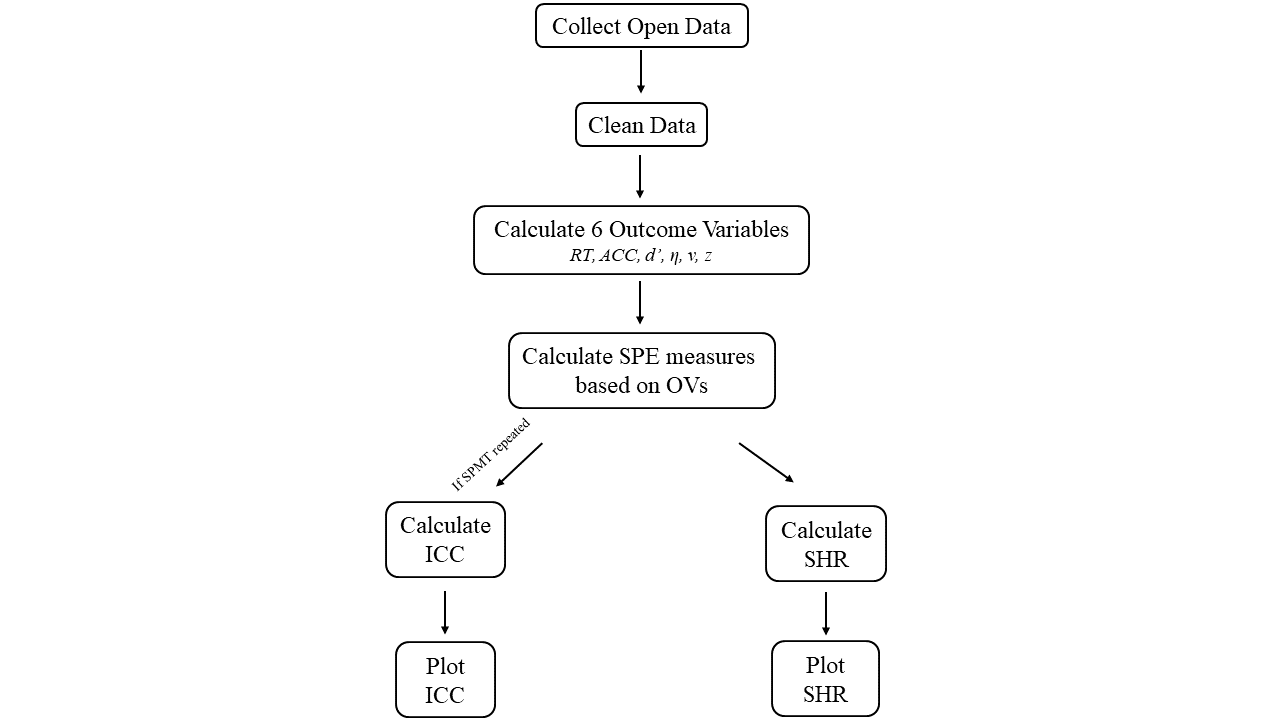
| Paper | Exp. | Independent Variable | | | | Sample  Size | # of Trials per Condition | Self-Prioritization Effect Indices | | | | | | Reliability | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| IV 1 | IV 2 | IV 3 | IV 4 | RT | ACC | d | Eff | v | z | ICC | SHR |
| Hu et al. (2023) | 1 | Matching | Identity | Emotion  **Control**, Neutral,  Happy, Sad | Session | 33 | 60 | √ | √ | √ | √ | √ | √ | √ | √ |
| Constable and Knoblich (2020) | 1 | Matching | Identity | Switch Identity  Partner, Stranger | Phase  **1**; 2 | 46 | 20 | √ | √ | √ | √ | √ | √ |  | √ |
| Constable et al. (2021) | 2 | Matching | Identity  Self; Stranger |  |  | 56 | 48 | √ | √ | √ | √ | √ | √ |  | √ |
| Qian et al. (2020) | 1 | Matching | Identity Self; Celebrity; Stranger | Mood (Session) |  | 24 | 24 | √ | √ | √ | √ | √ | √ |  | √ |
| 2 | Matching | Identity Self; Celebrity | Cue  With, **Without** |  | 25 | 25 | √ | √ | √ | √ | √ | √ |  | √ |
| Schäfer and Frings (2019) | 1 | Matching | Identity Self; Mother; Acquaintance/none |  |  | 32 | 18 | √ | √ | √ | √ | √ | √ |  | √ |
| 3 | 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 | Feedback | Sex | 105 | 60 | √ | √ | √ | √ | √ | √ |  | √ |
| Woźniak et al. (2018) | 1 | Matching | Identity | Facial Gender  Mele; Female |  | 18 | 56 | √ | √ | √ | √ | √ | √ |  | √ |
| 2 | Matching | Identity | Facial Gender  Mele; Female |  | 18 | 60 | √ | √ | √ | √ | √ | √ |  | √ |
| Liu et al. (2023) | 1 | Matching | Identity  Self; Stranger |  |  | 298 | 16 | √ | √ | √ | √ | √ | √ |  | √ |

*Note*. SPE: 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.

# **3 Analysis**

Analysis plans for this study were preregistered on OSF (https://osf.io/6dghm/). The drift rate (*v*) and starting point (*z*) of the drift-diffusion model (DDM) was obtained in Python, using the “RWiener” package (Wabersich & Vandekerckhove, 2014). All the other analyses in this paper were performed using the statistical software R (R Core Team, 2023).

The research flow of the current study is visually represented in Fig. 2. After collecting the data from each independent study, we performed data cleaning and calculated the six outcome variables as well as the SPE measures computed using different baseline conditions and outcome variables. Finally, we calculated the split-half reliabilities of these SPE measures. If there were test-retest data, we would also calculate the test-retest reliability using the intraclass correlation coefficient (ICC2, ICC2k).



**Fig. 2** Roadmap of the current study.

## 3.1 Deviation from Preregistration

In our pre-registration plan, we intended to estimate the drift rate (*v*) and starting point (*z*) of the drift-diffusion model (DDM) using the “fit\_ezddm” function from the “hausekeep” package (Lin et al., 2020). This function was a wrapper for the EZ-DDM function (Wagenmakers et al., 2007). However, parameter recovery found that this algorithm produces distinct result with the widely adopted HDDM package (see Supplementary for details). Thus, we used “RWiener” package (Wabersich & Vandekerckhove, 2014)., which performed much better in parameter recovery. Nevertheless, results from ezDDM can be found in detail in the supplementary materials (see supplementary, figure 1).

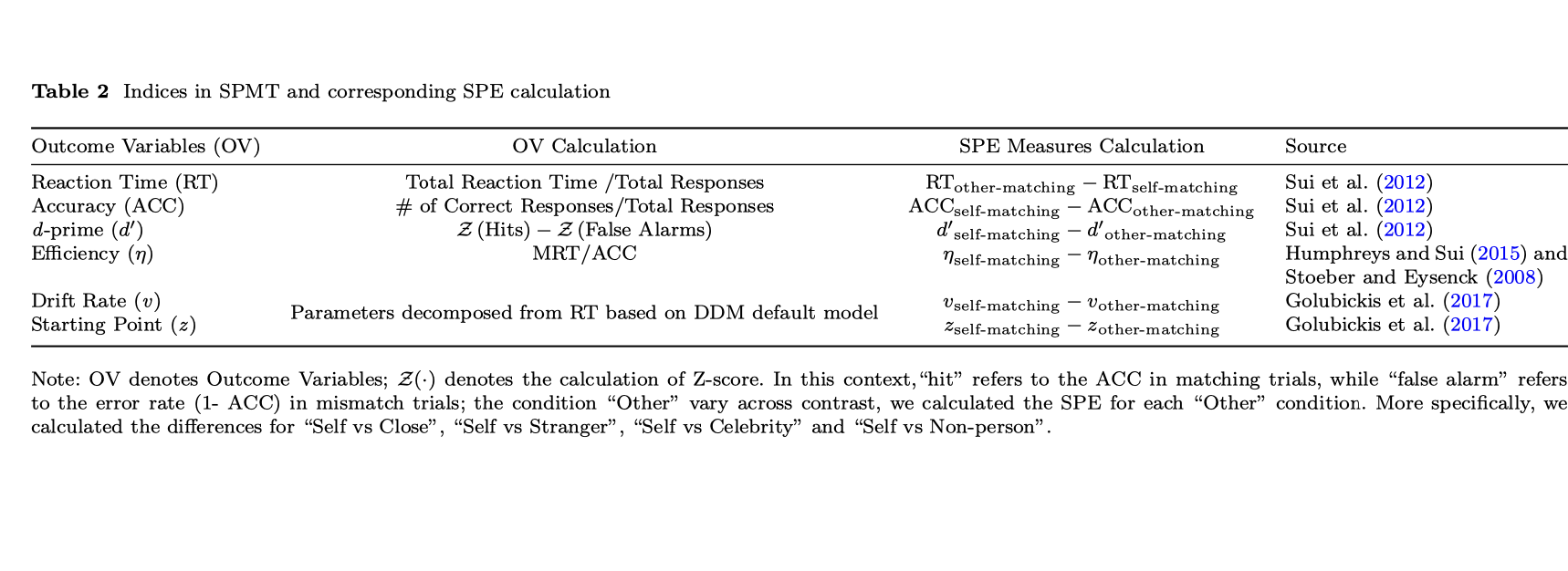
## 3.2 Data Pre-processing

In total, we gathered 18 publicly available datasets from 9 papers and 2 projects, as mentioned earlier and presented in Table 1. We used the following exclusion criteria for data preprocessing:

1. Participant exclusion criteria
2. Participants who had wrong trial numbers because of procedure errors is excluded from the analysis,
3. Participants with an overall accuracy < 0.5 is excluded from the analysis,
4. Participants with any of the conditions with zero accuracy is excluded from the analysis.
5. Trial level data exclusion criteria
6. Trials with no response or wrong key press is excluded from the analysis,
7. The practice trials are excluded from the formal analysis,
8. Participants with any of the conditions with zero accuracy is excluded from the analysis,
9. The data under conditions other than the “control condition” would not be used in the current study.

## 3.3 Calculating the Outcome Variables and SPE Measures

For each study, we calculated six outcome variables for each experimental condition: Mean Reaction Time (RT), Accuracy (ACC), *d* prime (*d′*), Efficiency (*η*), Drift Rate (*v*), and Starting Point (*z*). RT and ACC are obtained directly from the datasets, while *d′* and *η* are calculated based on RT and ACC (see Table 2). Please note that the condition “Other” vary across contrast, we calculated the SPE for each “Other” condition. More specifically, we calculated the differences for “Self vs Close other”, “Self vs Stranger”, “Self vs Celebrity” and “Self vs Non-person”.



## 3.4 Estimating the Reliability

**Weighted Average Split-half Reliability.** To ensure methodological rigor, we assessed the weighted average split-half reliability of the SPE measures using four distinct data splitting approaches: first-second, odd-even, permutated, and Monte Carlo (Kahveci et al., 2022; Pronk et al., 2022). The first-second split-half reliability involved dividing the trials into two halves: the first half and the second half. The odd-even split-half reliability separated the trials into sequences based on their odd or even numbers. On the other hand, the permutation split-half reliability shuffled the trial order and randomly assigned each half to a group. The Monte Carlo split-half approach was similar to the permutated split-half method, but it repeated the process thousands of times to calculate the average and 95% confidence interval of the split-half reliability. This comprehensive evaluation ensured the robustness of our findings.

First, the data were stratified per condition (Pronk et al., 2022) and then split into two halves. For example, when using Monte Carlo Split-Half, we randomly split the stratified data into two halves for 5000 times, which resulted in 5000 pairs of two halves of the data. Next, we calculated 5000 Pearson correlation coefficients based on these 5000 pairs of data and calculated the mean and 95% confidence interval of the 5000 correlations coefficients. The first-second split, odd-even split, and permutated split are similar to Monte Carlo method, but each only resulted in single reliability coefficient.

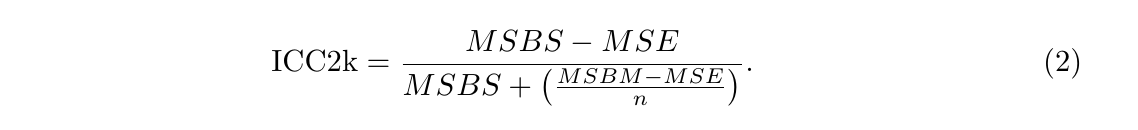
Finally, after computing split-half reliability coefficients for each dataset, substantial variations were observed in the split-half reliability across the datasets. To derive a representative overall value for each SPE measures, we adopted an approach akin to meta-analysis, allowing for a more accurate estimation of the average split-half reliability. As noted by Kucina et al. (2023), the number of trials significantly influences the reliability of cognitive experiments, with higher trial numbers resulting in increased reliability (similar pattern was also found in SPMT, see Supplementary Fig. 6). Hence, to account for the impact of trials, we assigned weights to the split-half reliability based on the trial numbers present in each dataset using the “aggregate.escalc” function from the “metafor” Package (Viechtbauer, 2010). Employing a method similar to weighted averaging, we obtained the final weighted average split-half reliability (WASR) as well as its 95% confidence interval, ensuring a comprehensive and robust evaluation of the SPE measures across the datasets.

**Test-Retest Reliability (ICC).** We assessed the test-retest reliability of the six outcome variables for our dataset that involved test-retest sessions by calculating the Intraclass Correlation Coefficient (ICC) using “psych” package (Revelle, 2017). ICC is a well-established measure used in test-retest, intra-rater, and inter-rater studies to assess reliability (Fisher, 1992). Unlike the Pearson correlation coefficient, ICC considers both the correlation and agreement between multiple measurements, making it a more comprehensive measure of test-retest reliability.

Our primary aim is to evaluate the appropriateness of the SPMT in assessing individual differences and group-level variations. To achieve this objective, we focused on using two specific types of Intraclass Correlation Coefficients (ICC) within the ICC family, namely ICC2 and ICC2k. ICC2 focuses on the individual-level reliability of the indices, while ICC2k evaluates the reliability of mean ratings furnished by a group of judges (Koo & Li, 2016). For the calculation of ICC2 estimates, the formula is:A math equations on a white background

Description automatically generated

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. For the calculation of ICC2k estimates, the formula is:



Although there is no strict criterion for defining the level of reliability, a widely accepted guideline for Cronbach’s alpha is that a value of 0.60 is “acceptable”, and a value greater than 0.8 means excellent reliability (Cicchetti & Sparrow, 1981; Kupper & Hafner, 1989).

# **4** **Results**

In 18 independent studies, 14 studies have data for the baseline condition of Self vs Close other, 12 studies have the data for Self vs Stranger, 2 studies have the data for Self vs Celebrities, 1 study has the data for Self vs none condition. We reported the results for all these baseline conditions.

## 4.1 Split-Half Reliability

As described in method part, we utilized four different methods to calculate split-half reliability, namely the first-second, odd-even, permuted, and Monte Carlo methods. The paper exclusively presents the results obtained from the Monte Carlo split-half method due to its robustness, as highlighted in Pronk et al. (2022). The results of the other three split-half methods can be found in the supplementary materials (see Supplementary Fig. 1 & 2), which also have quite similar result pattern with the Monte Carlo methods.

The weighted average split-half reliabilities (WASR) using of Monte-Carlo method for different SPE measures is presented in figure 3. Among the measured outcome variables, the four highest ranking weighted average split-half reliabilities (WASR) are as follows: Reaction Time (RT) with the "Stranger" contrast (WASR = .65, SE = .02, p <.001, 95% CI [.60, .70]); Efficiency (*η*) with the "Stranger" contrast (WASR = .63, SE = .02, p <.001, 95% CI [.58, .68]); RT with the "Close other" contrast (WASR = .58, SE = .02, p <.001, 95% CI [.54, .63]); *η* with the "Self vs Close" contrast (WASR = .57, SE = .02, p <.001, 95% CI [.52, . 61]). These SPE measures achieved a split-half reliability around 0.6 or higher, which is considered an acceptable level of reliability.

For the remaining 4 outcome variables, the reliability is around 0.5 or lower, indicating poor test-retest reliability. These include ACC, *d’*, *v* and *z*. It is notice that almost all the WASR of *z* using different contrasts are around 0, which can be considered to have no reliability.

图表, 条形图

描述已自动生成

**Fig. 3.** The weighted average of Monte-Carlo split-half reliabilities for different SPE measures.

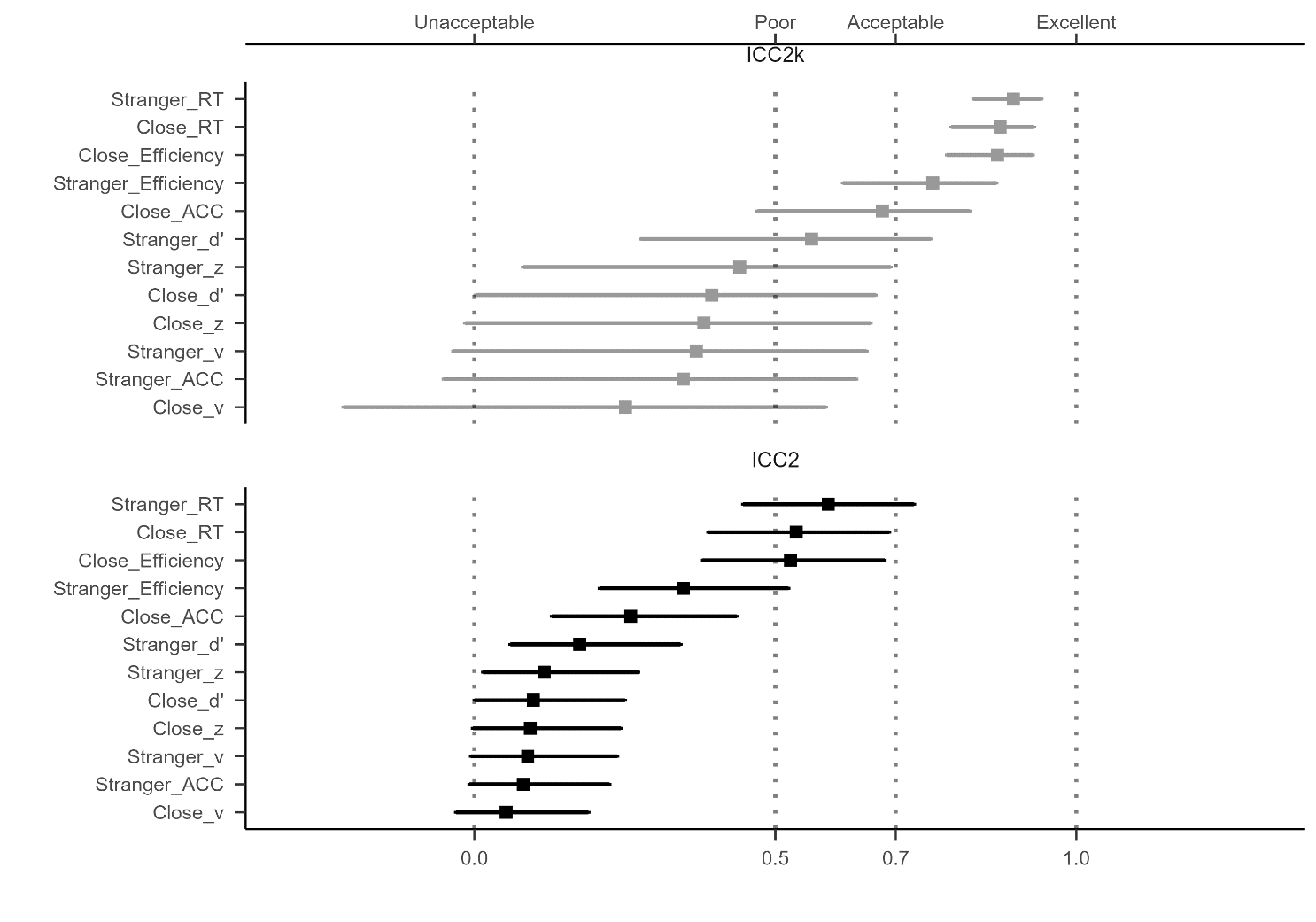
Note: The vertical axis represents 24 different SPE measures, combining 6 outcomes variables (RT, ACC, *d’*, *η*, *v*, *z*) and 4 contrasts (close other, stranger, celebrity, and non-person). The weighted average split-half reliability and 95% confidence intervals are shown by points and lines.

## 4.2 Intraclass correlation coefficient (ICC)

It is important to note that we could only calculate ICC for the study in our labprtory (Hu et al., 2023) since all other datasets did not include re-test sessions. For our own dataset, we only have 2 contrasts, the “Close other” and “Stranger”. To test the robustness of the results reported here, we also explored two datasets that included re-test session but devivated quite strongly from the original SPMT (see Supplementary Fig. 5).

As illustrated in Fig. 4, the ICC2k values for Reaction Time (RT) and Efficiency demonstrate notably high levels of reliability, while the other outcome variables exhibit unacceptable to poor ICC2k. When considering the "Close" contrast, the ICC2k for RT is 0.87 (95% CI = [.79, .93]), and for Efficiency, it is 0.86 (95% CI = [.78, .93]). Similarly, for the "Stranger" contrast, the ICC2k for RT is 0.89 (95% CI = [.82, .94]), and for Efficiency, it is 0.76 (95% CI = [.61, .87]).

However, it is worth noting that RT and Efficiency display ICC2 values ranging from poor to acceptable. Specifically, under the "Close" contrast, the ICC2 for RT is 0.53 (95% CI = [.39, .69]), and for Efficiency, it is 0.52 (95% CI = [.38, .68]). Meanwhile, for the "Stranger" contrast, the ICC2 for RT is 0.58 (95% CI = [.45, .73]), and for Efficiency, it is 0.34 (95% CI = [.21, .52]). All the other outcome variables exhibit unacceptable to poor ICC2.



**Fig. 4** Intraclass correlation coefficient.

*Note:* The vertical axis represents 12 different SPE measures, combining 6 outcomes variables (RT, ACC, d’, η, v, z) and 2 contrasts (close other, stranger).The upper facet of the figure presents the results for ICC2k, while the lower facet displays the results for ICC. The ICC values and their corresponding 95% confidence intervals are illustrated using points and lines.

# **5 Discussion**

Despite the significance of assessing the reliability of a behavioral paradigm before its implementation, this practice is not yet extensively embraced by researchers (Green et al., 2016; Hedge et al., 2018; Parsons et al., 2019). In this pre-registered study, our objective is to investigate the reliability of the outcome variables related to the SPE measures in the SPMT. To achieve this, we conducted a re-analysis of data from 18 datasets across 9 articles and 2 projects, utilizing split-half reliability and intraclass correlation coefficient (ICC2, ICC2k). Our findings reveal that the reaction time and efficiency demonstrated better results in terms of split-half reliability and test-retest reliability (ICC2, ICC2k), with weight average split-half reliability around 0.6 or higher, ICC2 around 0.5 or higher, ICC2k higher than 0.7. All the other outcome variables performed poorly, with weight average split-half reliability lower than 0.5, ICC2 lower than 0.5, ICC2k lower than 0.7. The comprehensive analysis of these datasets collectively suggests that reaction time and efficiency are the most reliable outcome variables among all measured in the SPMT. Moreover, the pattern observed in the ICC values suggests that reaction time and efficiency in the SPMT are better suited for group-level analysis rather than assessing individual-level variation.

Although the Reaction Time (RT) and Efficiency indices yielded relatively better results compared to other indices, they still fell short of achieving commonly considered excellent reliability levels (typically a value greater than 0.8 means excellent reliability) (Cicchetti & Sparrow, 1981; Kupper & Hafner, 1989). The low split-half reliability and low ICC2 may suggest that the SPMT is not suitable for assessing individual-level variation. Several plausible reasons as well as potential solutions can be identified. Firstly, the insufficient number of trials per condition may contribute to the low split-half reliability. A recent study by Kucina et al. (2023) emphasized the importance of trial numbers for cognitive tasks in determining reliability. The findings revealed that increasing the number of trials and considering greater conflict effects or individual differences can enhance reliability compared to the original paradigm. Specifically, the study identified that satisfactory reliability required 48 or more trials, while achieving a higher level of reliability necessitated 72 trials. As shown in Supplementary (Fig. 6), we calculated the correlation coefficient between the trial numbers and the Monte Carlo split-half reliabilities. We found a strong positive correlation between several metrics, such as RT and efficiency. Therefore, incorporating a higher number of trials in future implementations of the SPMT paradigm may enhance the split-half reliability by improving measurement consistency. Secondly, the presence of a practice effect or fatigue effect could be another significant factor influencing reliability. If these effects are substantial enough to cause a noticeable change in participants' performance between measurement occasions, it can introduce additional variability in the measurements and lower the ICC2 values (Oswald et al., 2015; Siegelman et al., 2017). This highlights the need for a more nuanced task setting that can consistently capture performance nuances and reveal individual differences more sensitively (Hedge et al., 2018). Recent study on gamification of cognitive tasks have shown that incorporating gamification elements can effectively improve data quality and assessment efficacy (Friehs et al., 2020). Therefore, modifying the SPMT with a more dynamic paradigm, such as incorporating gamification elements, may enhance split-half reliability by improving measurement consistency. Finally, it is worth mentioning the influence of serial dependence effects on task reliability. A recent set of studies has examined serial dependence effects in a variety of cognitive tasks (Braun et al., 2018; Zhang & Alais, 2020). Serial dependence refers to the phenomenon in which the outcome of one trial is influenced by preceding trials, resulting in a systematic relationship between consecutive trials (Pascucci et al., 2023). Notably, studies in the field of perceptual decision making have demonstrated strong serial dependence effects in perception, even when the visual stimuli were reliable and varied randomly over time (Fischer & Whitney, 2014; John-Saaltink et al., 2016). In particular, if the split-half design unintentionally separates temporally adjacent trials in the SPMT, the presence of serial dependence may introduce performance differences between the halves, leading to a reduction in the reliability estimate. Thus, to accurately control for the impact of serial dependence in experiments, further research should employ appropriate statistical methods that account for the temporal dependencies between trials. Time series analysis techniques (Huitema, 1986) or modeling approaches that capture the serial correlation (Mei et al., 2023) can be utilized to obtain more accurate results.

We would like to emphasize the importance of considering the parameters (*v* and *z*) derived from the drift diffusion model within the SPMT. Our findings indicate that the weighted average split-half reliability and test-retest reliability of the model parameters was notably low (WASR lower than 0.5 for *v*, around 0 for *z;* ICCs lower than 0.5). This raises concerns about the suitability of applying the default drift diffusion model in cognitive tasks like SPMT. Since the default drift diffusion model relies on specific assumptions regarding the cognitive processes underlying reaction time (specifies fixed values for drift, bound height or noise level, starting position bias, and non-decision time) (Ratcliff, 1978; Ratcliff et al., 2016) , if these assumptions fail to accurately capture the true underlying mechanisms or if they are violated within the assumption of SPMT, it can result in inconsistent parameter estimates and diminished the reliability (Johnson et al., 2017). Thus, our results raise significant questions regarding the efficacy of applying the classical hierarchical drift diffusion model to the SPMT. Further research is needed to comprehensively investigate the underlying causes of low split-half reliability and test-retest reliability. Additionally, it is important to develop extensions to the drift diffusion model and explore alternative modeling approaches that can produce more reliable parameter estimates. Due to the methodological challenges involved in introducing innovations to the model, several researchers have proposed user-friendly frameworks for building and fitting extensions to the drift diffusion model. Examples include the generalized drift-diffusion model (GDDM) by Shinn et al. (2020), RLDDM modules in HDDM by Pedersen and Frank (2020). These extensions offer alternative modeling approaches that could be used to better capture the true underlying mechanisms of reaction time in SPMT. These efforts could contribute to the refinement and optimization of the model for future research applications, enabling more reliable and valid interpretations of cognitive processes in the SPMT and similar paradigms.

It is important to note that ICC values should not be interpreted solely as a measure of the test's overall quality, but rather as an indication of the specific types of questions it can effectively address (Koo & Li, 2016). The results obtained from the intraclass correlation coefficients indicate that the SPMT, particularly the Response Time and efficiency, is more appropriate for group-level analysis rather than assessing individual-level variation. Specifically, we found that the RT and efficiency measures exhibited excellent test-retest reliability at the group level (ICC2k), regardless of whether the target is, indicating strong group-level consistency over time. On the other hand, the ICC2 values for RT and Efficiency were relatively lower, below 0.7. This discrepancy suggests that these outcome variables are more influenced by variations between participants rather than within participants. Similar results patterns have been observed in previous research examining other cognitive paradigms such as Flanker, Simon, or Stroop tasks (Clark et al., 2022; Mollon et al., 2017). These findings also align with the concept of the reliability paradox proposed previously (Hedge et al., 2018; Logie et al., 1996). Behavioral paradigms, including the SPMT, are susceptible to factors such as external conditions and contextual differences, which contribute to greater within-participant variability and lower ICC2 values. However, when averaging performance across different individuals, the task still demonstrates good consistency, leading to higher ICC2k values (Liljequist et al., 2019). In practical terms, these results suggest that the SPMT is better suited for discerning performance differences between individuals or groups rather than capturing consistent performance within the same individuals over time. Therefore, researchers should consider these factors when investigating individual differences using the SPMT.

Our study has several limitations that should be acknowledged. Firstly, although we made efforts to enhance sample diversity by including open data whenever possible, it is important to note that the majority of our samples still consisted of individuals from what is commonly referred to as "wired" populations (Rad et al., 2018). 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. Additionally, when assessing the intraclass correlation coefficients (ICCs), only our own dataset had longitudinal data available, which could potentially limit the representativeness of the results. Furthermore, the majority of the studies included in our analysis focused on adults from healthy populations. Therefore, further investigation is needed to include more datasets with a more diverse population in order to determine the reliability of the SPMT in different settings. Finally, it is important to clarify the aim of our study, which primarily focused on exploratory purposes and providing information regarding the current state of reliability for the assessed outcome variables. Consequently, future research could concentrate on modifying the paradigm and conducting tests to assess potential improvements. We propose several approaches that could be considered, such as introducing more challenging task variations (gamification) that have the potential to increase the reliability of accuracy measurements. Another suggestion is to include a greater number of trials for each condition, as this may contribute to improved reliability. It is strongly encouraged to undertake further investigation and experimentation in order to refine the paradigm and enhance the reliability of the indices, rather than dismissing the paradigm under certain circumstances.

In conclusion, this study provides significant insights into the reliability of the self-perceptual matching task (SPMT) and highlights important considerations for interpreting its reliabilities. We have demonstrated that the Response Time and Efficiency measures exhibit greater reliability compared to other outcome variables in the SPMT. Furthermore, our findings indicate that the SPMT is more suitable for group-level analysis rather than assessing individual-level variation. These findings have important implications for future task design and data collection protocols aimed at improving reliability. Ultimately, our study paves the way for the prospective utilization of these tasks, in various domains including research, clinical applications, and personal performance monitoring. The information obtained from our study contributes valuable knowledge to the field and sets the stage for further investigations and advancements in utilizing SPMT effectively.

# **Acknowledgements**

The present research is support by.

# **Author contributions**

HCP contributed to the conception and supervision of the study. JS contributed to fund raising, HCP contributed to data collection. ZL, ZYR write the simulation code for pre-registration. HMZ collected the datasets and performed the data pre-processing, analysis and visualize the results. In addition, ZL, HMZ and HCP contribute to discussing the results and the drafting of the final manuscript. All authors critically revise the manuscript.

# **Data and Material Availability**

The pre-registration plan is available at <https://osf.io/zv628>. The de-identified raw data from our lab is available at <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 at <https://github.com/Chuan-Peng-Lab/ReliabilitySPE>.

# **Competing interests**

The authors declare no competing interests.

**Supplementary Information**

## 1.1 Parameter Recovery Result for Package Comparison

We conducted a package comparison by generating 100 datasets using the DDM package HDDM in Python. ~~(The computation time required by HDDM is excessively long, and its accuracy is not significantly higher than RWiener. Therefore, we ultimately decided to abandon the use of HDDM as our computational tool.)~~ These datasets were specifically configured with parameters a=2, t=0.3, v=1, and z=0.7. Subsequently, we utilized three different DDM packages in R (RWiener, hausekeep, and FastDMinR) to compute parameter estimates for these generated datasets. The evaluation process involves comparing the computed values obtained from the R packages with the set parameters (a=2, t=0.3, v=1, and z=0.7). If the computed values from the R packages are found to be closer to the set values, it signifies that the respective R package provides more accurate parameter estimation for the drift-diffusion model.

The results of package comparison are shown in figure 6,

The estimate of v obtained from RWiener is 1.01, 95% CI [.98, 1.03], and the estimate of z is 0.77, 95% CI [.76, .78], both of which are very close to the values we pre-defined. Considering the parameters calculated using other packages are either highly inaccurate, have excessively wide confidence intervals, or require very long computation times, we have decided to use RWiener for calculating our results. It provides a good balance between accuracy, confidence interval width, and computational efficiency, making it the most suitable choice for our analysis.

图形用户界面, 图示, 应用程序

描述已自动生成

**Supplementary Fig. 6 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.

## WASR for SPE Measures Using other Three Split-half Methods

In this section, we present the WASR result for SPE measures using other three split-half methods (first-second, odd-even, permutated).

**电脑屏幕截图

描述已自动生成**

**Supplementary Fig. 2** WASR for SPE measures using other three split-half methods.

*Note:* The vertical axis of the graph illustrates eight distinct outcome variables, which includes two additional indices from the DDM, implemented using the "hausekeep" package. The line and dots on the graph represent the weighted average split-half reliability values along with their corresponding 95% confidence intervals. In cases where only one point is visible, it signifies that the confidence interval for that specific point estimate extends beyond the range of our coordinate axes (0, 1). The figure is divided into separate facets arranged from left to right, each representing WASR calculated using three distinct methods: first-second, odd-even, and permuted. Moving from top to bottom, each facet in the graph corresponds to a different target for the SPE measures, specifically friend, stranger, celebrity, and non-person.

## SHR for SPE Measures Using Four Split-half Methods

In this section, we present the results of Monte Carlo SHR (figure 3) and SHR obtained using the other three split-half methods (figure 4) for estimating the drift rate (v) and starting point (z) from the ezDDM ("hausekeep" package). It should be noted that the estimation of parameter "a" in hausekeep is highly deviant from HDDM due to the assumption of z = a / 2 (see supple figure 1). Consequently, we do not report the results obtained from this package in the main text.

图片包含 图示

描述已自动生成

**Supplementary Fig. 3.** Monte Carlo SHR for SPE measures.

*Note:* The vertical axis of the graph illustrates eight distinct outcome variables, which includes two additional indices from the DDM, implemented using the "hausekeep" package. The line and dots on the graph represent the split-half reliability estimated using Monte Carlo split-half method, along with their corresponding 95% confidence intervals. If only the point is shown in the graph, it indicates that the confidence interval for that point estimate extends beyond the range of our coordinate axes (0, 1).

图示

描述已自动生成

**Supplementary Fig. 4 SHR for SPE measures using other three split-half methods.**

*Note:* The vertical axis of the graph illustrates eight distinct outcome variables, which includes two additional indices from the DDM, implemented using the "hausekeep" package. The figure is divided into separate facets arranged from left to right, each representing SHR calculated using three distinct methods: first-second, odd-even, and permuted. Moving from top to bottom, each facet in the graph corresponds to a different target for the SPE measures, specifically friend, stranger, celebrity, and non-person.

From the result, it is evident that ezDDM demonstrates higher stability in estimating the drift rate (v). This may be attributed to the estimation method used in "hausekeep," which relies on average reaction time and accuracy, as opposed to the method employed in "RWiener," which relies on individual trial-level reaction times and correctness. The chosen split-half procedure may have a greater impact on the estimation by "RWiener," leading to lower split-half reliabilities for both of its indices. These observations imply that the reliability of calculating DDM parameters through split-half reliabilities is influenced by two key factors. Firstly, the stability of the parameter itself plays a significant role, and secondly, the stability of the package used to compute that parameter is equally crucial.

## ICCs for different SPE indices.

电脑萤幕画面

描述已自动生成

**Supplementary Fig. 5** ICCs for different SPE indices.

*Note:*The vertical axis of the graph illustrates eight distinct outcome variables, which includes two additional indices from the DDM, implemented using the "hausekeep" package. The line and dots on the graph represent the value of ICC, along with their corresponding 95% confidence intervals. If only the point is shown in the graph, it indicates that the confidence interval for that point estimate extends beyond the range of our coordinate axes (0, 1). The upper facet of the figure presents the results for ICC2k, while the lower facet displays the results for ICC.

## Correlation Between Monte Carlo SHR and Trial Numbers

In this section, we present the results of the correlation between Monte Carlo SHR and trial numbers (figure 6). Some outcome variables, such as RT, efficiency shown significant correlation between Monte Carlo SHR and trial numbers. However, for outcome variables like *d’* and *v*, there is minimal correlation between trial numbers and SHR. It is found that the SPMT paradigm requires approximately 80 trials to achieve a Monte Carlo SHR of 0.8 when the SPE measure is RT under 'Stranger', and about 120 trials under 'Close.' Additionally, achieving a Monte Carlo split-half reliability of 0.8 for *v* may require more than 130 trials. On the other hand, high Monte Carlo SHR is challenging to achieve for the remaining three outcome variables, particularly for the *z*, even with 150 or more trials.

It is important to emphasize that here we only conducted a simple regression analysis of trial numbers and Monte Carlo SHR based on the collected datasets. 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. However, considering the significant correlation between the number of trials and Monte Carlo split-half reliability, our findings suggest that, for the SPMT paradigm, achieving higher reliability for clinical evaluation would necessitate conducting more trials. As a reasonable approach, it might be beneficial to consider having more than 120 trials under each experimental condition.

图表

描述已自动生成

**Supplementary Fig. 6 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.

**References**

Braun, A., Urai, A. E., & Donner, T. H. (2018). Adaptive history biases result from confidence-weighted accumulation of past choices. *Journal of Neuroscience, 38*(10), 2418-2429.

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>

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.

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>

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>

Conway, M. A., & Dewhurst, S. A. (1995). The self and recollective experience. *Applied Cognitive Psychology, 9*(1), 1-19. <https://doi.org/10.1002/acp.2350090102>

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268-294. <https://doi.org/10.1037/0096-3445.104.3.268>

Cunningham, S. J., & Turk, D. J. (2017, Jun). Editorial: A review of self-processing biases in cognition. *Quarterly journal of experimental psychology, 70*(6), 987-995. <https://doi.org/10.1080/17470218.2016.1276609>

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>

Desebrock, C., Sui, J., & Spence, C. (2018). Self-reference in action: Arm-movement responses are enhanced in perceptual matching. *Acta psychologica, 190*, 258-266. <https://doi.org/10.1016/j.actpsy.2018.08.009>

Enock, F., 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>

Enock, F. E., Hewstone, M. R., Lockwood, P. L., & Sui, J. (2020). Overlap in processing advantages for minimal ingroups and the self. *Scientific Reports, 10*(1), 18933. <https://doi.org/10.1038/s41598-020-76001-9>

Feng, C., Yan, X., Huang, W., Han, S., & Ma, Y. (2018). Neural representations of the multidimensional self in the cortical midline structures. *NeuroImage, 183*, 291-299. <https://doi.org/10.1016/j.neuroimage.2018.08.018>

Feng, C., Zhou, X., Zhu, X., Zhu, R., Han, S., & Luo, Y.-J. (2020). Effect of intranasal oxytocin administration on self-other distinction: Modulations by psychological distance and gender. *Psychoneuroendocrinology, 120*, 104804. <https://doi.org/10.1016/j.psyneuen.2020.104804>

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature neuroscience, 17*(5), 738-743.

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.

Gillespie‐Smith, K., Ballantyne, C., Branigan, H. P., Turk, D. J., & Cunningham, S. J. (2018). The I in autism: Severity and social functioning in autism are related to self‐processing. *British journal of developmental psychology, 36*(1), 127-141. <https://doi.org/10.1111/bjdp.12219>

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., Falben, 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. *Mem Cognit, 45*(7), 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.

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.

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>

Hu, C.-P., Peng, K., & Sui, J. (2023, 2023-05-06). Data for Training Effect of Self Prioritization[DS/OL]. V1. *Science Data Bank*. <https://doi.org/10.57760/sciencedb.08117>.

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>

Huitema, B. E. (1986). Autocorrelation in behavioral research: Wherefore art thou? In *Research methods in applied behavior analysis: Issues and advances* (pp. 187-208). Springer.

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>

Jiang, M., Wong, S. K. M., Chung, H. K. S., Sun, Y., Hsiao, J. H., Sui, J., & Humphreys, G. W. (2019). Cultural Orientation of Self-Bias in Perceptual Matching. *Front Psychol, 10*, 1469. <https://doi.org/10.3389/fpsyg.2019.01469>

John-Saaltink, E. S., Kok, P., Lau, H. C., & De Lange, F. P. (2016). Serial dependence in perceptual decisions is reflected in activity patterns in primary visual cortex. *Journal of Neuroscience, 36*(23), 6186-6192.

Johnson, D. J., Hopwood, C. J., Cesario, J., & Pleskac, T. J. (2017). Advancing Research on Cognitive Processes in Social and Personality Psychology:A Hierarchical Drift Diffusion Model Primer. *Social Psychological and Personality Science, 8*(4), 413-423. <https://doi.org/10.1177/1948550617703174>

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>

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., ..., & 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>

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.

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.

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

Liu, Y. S., 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>

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.

Maire, H., Brochard, R., & Zagar, D. (2020). A Developmental Study of the Self‐Prioritization Effect in Children Between 6 and 10 Years of Age. *Child development, 91*(3), 694-704. <https://doi.org/10.1111/cdev.13352>

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>

Mei, N., Rahnev, D., & Soto, D. (2023). Using serial dependence to predict confidence across observers and cognitive domains. *Psychonomic Bulletin & Review*, 1-13.

Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences in visual science: What can be learned and what is good experimental practice? *Vision research, 141*, 4-15.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly journal of experimental psychology, 11*(1), 56-60. <https://doi.org/10.1080/17470215908416289>

Navon, M., & Makovski, T. (2021). Are Self-related Items Unique? the Self-prioritization Effect Revisited. <https://doi.org/10.31234/osf.io/9dzm4>

Nijhof, A. D., & Bird, G. (2019). Self‐processing in individuals with autism spectrum disorder. *Autism research, 12*(11), 1580-1584. <https://doi.org/10.1002/aur.2200>

Orellana-Corrales, G., Matschke, C., & Wesslein, A.-K. Does Self-Associating a Geometric Shape Immediately Cause Attentional Prioritization? Comparing Familiar vs. Recently Self-Associated Stimuli in the Dot-Probe Task. <https://doi.org/10.1027/1618-3169/a000502>

Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods, 47*, 1343-1355.

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.

Pascucci, D., Tanrikulu, Ö. D., Ozkirli, A., Houborg, C., Ceylan, G., Zerr, P., Rafiei, M., & Kristjánsson, Á. (2023). Serial dependence in visual perception: A review. *Journal of Vision, 23*(1), 9-9.

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>

Pedersen, M. L., & Frank, M. J. (2020). Simultaneous hierarchical bayesian parameter estimation for reinforcement learning and drift diffusion models: a tutorial and links to neural data. *Computational Brain & Behavior, 3*, 458-471.

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. (2023). R: A Language and Environment for Statistical Computing. <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.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review, 85*(2), 59.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in cognitive sciences, 20*(4), 260-281.

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, Sep). Self-reference and the encoding of personal information. *J Pers Soc Psychol, 35*(9), 677-688. <https://doi.org/10.1037//0022-3514.35.9.677>

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>

Shapiro, K. L., Caldwell, J., & Sorensen, R. E. (1997). Personal names and the attentional blink: a visual "cocktail party" effect. *J Exp Psychol Hum Percept Perform, 23*(2), 504-514. <https://doi.org/10.1037//0096-1523.23.2.504>

Shinn, M., Lam, N. H., & Murray, J. D. (2020). A flexible framework for simulating and fitting generalized drift-diffusion models. *ELife, 9*, e56938.

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods, 49*, 418-432.

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>

Strachan, J. W., Constable, M. D., & Knoblich, G. (2020). It goes with the territory: Ownership across spatial boundaries. *Journal of Experimental Psychology: Human Perception and Performance, 46*(8), 789. <https://doi.org/10.1037/xhp0000742>

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. (2013, Nov). Self-referential processing is distinct from semantic elaboration: Evidence from long-term memory effects in a patient with amnesia and semantic impairments. *Neuropsychologia, 51*(13), 2663-2673. <https://doi.org/10.1016/j.neuropsychologia.2013.07.025>

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>

Symons, C. S., & Johnson, B. T. (1997, May). The self-reference effect in memory: a meta-analysis. *Psychological Bulletin, 121*(3), 371-394. <https://doi.org/10.1037/0033-2909.121.3.371>

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.

Wabersich, D., & Vandekerckhove, J. (2014). The RWiener Package: an R Package Providing Distribution Functions for the Wiener Diffusion Model. *R Journal, 6*(1).

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>

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, Y., 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>

Zhang, H., & Alais, D. (2020). Individual difference in serial dependence results from opposite influences of perceptual choices and motor responses. *Journal of Vision, 20*(8), 2-2.

Zhou, A., Duan, B., Wen, M., Wu, W., Li, M., Ma, X., & Tan, Y. (2019). Self-referential processing can modulate visual spatial attention deficits in children with dyslexia. *Frontiers in Psychology, 10*, 2270. <https://doi.org/10.3389/fpsyg.2019.02270>

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>