Reliability of Self-Prioritization Effect as Measured by the Self-Perceptual Matching Task: Evidence from Multiple Datasets

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

[1*]School of Psychology, Nanjing Normal University, Nanjing, China.
[2*]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-prioritization effect (SPE) refers to the effect that performance on cognitive tasks is better when stimuli are related to the self than when they are not. In the last decade, the self -perceptual matching task (SPMT) has emerged as a mainstream paradigm for studying SPE due to its simplicity and elimination of familiarity effects. As a simple button-pressing task, SPMT yields two outcomes: reaction time and accuracy. Other indices can be derived from reaction times and accuracy, including sensitivity d-prime under the signal-detection theory, the efficiency index from the direct division of RT and ACC, and drift rate (v) and starting point (z) estimated using the drift-diffusion models. All these indices have been used to quantify SPE in the literature. However, the reliability of these SPE indices remains unexplored. To address this research gap, we performed a pre-registered study in which we re-evaluated data from 18 datasets across 11 articles using the intraclass correlation coefficient (ICC) and split-half reliability. Our results reveal that response time exhibits high and consistent test-retest reliability across datasets, while accuracy-based measurements yield sub-optimal outcomes. The ICC results suggest that all the indices related to SPE in the SPMT are more suitable for group-level analysis rather than assessing individual-level variation. These findings establish a benchmark for future investigations utilizing the SPMT and underscore the limitations of accuracy-based measures, which should be considered when employing the SPMT as an assessment tool.

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 effect 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 then 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 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 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) proposed a paradigm where participants first associate geometrical shapes (e.g., triangle, square, and circle) with labels of persons (e.g.,“You,” “friend,” and“stranger”) and then perform a perceptual matching task in which they decide if the shape-label pairs presented on the screen match the learned association or not (Sui et al., 2012). Because the task requires participants to learn the social meaning of different geometric shapes, it is called the Self-Perceptual Matching Task (SPMT). In this task, Sui et al. (2012) found that shapes associated with the self are performed better, with faster response times, better accuracy, and/or higher sensitivity scores, compared to shapes associated with friends and strangers. Because the self-relatedness is acquired immediately right before they start the perceptual matching task, this paradigm eliminated the effect of familiarity of the stimuli.

Since then, 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), as well as “good self” and “bad self” labels (Hu et al., 2020), and the group advantage effect of in-group labels (Constable, Elekes, et al., 2019; Constable & Knoblich, 2020; Enock et al., 2020; Enock et al., 2018). Moreover, the SPMT has been applied to various fields.

In neuroscience and physiology, researchers investigate which brain regions are activated during self-prioritization effect (Feng et al., 2018; Humphreys & Sui, 2015), and gender differences in self-prioritization effect due to oxytocin (Feng et al., 2020). In clinical research, SPMT has been used 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). Cross-cultural studies have shown that individuals from individualistic cultures demonstrate a stronger self-prioritization effect (Jiang et al., 2019), and that the language of the experimental stimuli can affect the strength of the effect (Ivaz et al., 2016). Finally, the SPMT has also been applied to child development, with studies examining developmental changes in self-positivity effects (Maire et al., 2020; Zhou et al., 2019).

While SPMT has gained widespread adoption as a prominent method for investigating the underlying mechanism of the self-prioritization effect, there has been microscopic examination and report of the psychometric properties of the outcomes, necessitating a careful evaluation (Parsons et al., 2019; Zorowitz & Niv, 2023). Given the increasing use of SPMT to assess individual differences in fields such as psychiatry (Y. S. 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 abilities (Parsons et al., 2019). Furthermore, in tasks as simple as the SPMT, there are multiple approaches to quantify the self-prioritization effect. These include two direct measures based on SPMT, namely reaction times (RT) and accuracy (ACC), as well as derived measures such as efficiency (Humphreys & Sui, 2015; Stoeber & Eysenck, 2008), d-prime of Signal Detection Theory (SDT) (Hu et al., 2020; Sui et al., 2012), and drift rate (v) and starting point (z) from Drift Diffusion Model (DDM) (Golubickis et al., 2017). Consequently, two important questions remain unanswered: (1) Do these indices reliably capture the self-prioritization effect across time points? and (2) If so, which index is most suitable for repeated measurements? Addressing these questions is crucial for establishing the reliability and validity of SPMT measurements, allowing for accurate assessment of the self-prioritization effect and its implications in various domains.

To address the existing research gap, the present study aimed to investigate the reliability of self-prioritization effect (SPE) indices in the self-perceptual matching task (SPMT). In order to comprehensively assess the SPE indices derived from SPMT, we examined six indices as mentioned earlier, that capture the disparity between self-related and other-related stimuli of the matching trials. This was achieved by reanalyzing data obtained from previous studies that employees SPMT. Given the diverse methods available for evaluating the reliability of cognitive tasks, we employed both the Split-Half Reliability and Intraclass Correlation Coefficient (ICC) to determine the reliability of each SPE index. These findings aim to provide valuable insights into the reliability and consistency of SPMT and its indices, having the potential to facilitate the future utilization of SPMT in research, clinical settings, and personal performance monitoring.

2 Methods

2.1 Ethics approval

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 and Datasets

In order to assess the reliability of SPMT, we first provided a brief overview of its original experimental design, 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 phases (see Fig. 1). In the first phase (learning phase), 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 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.

In this study, we collected a total of 18 existing datasets derived from 11 research articles, and one from our laboratory (Hu et al., 2023) and one from our collaborators (T. Liu et al., 2023), that included raw data from empirical studies utilizing the SPMT. The selection of these datasets was based on two criteria: (1) the experimental design did not deviate from the original SPMT of Sui et al. (2012); (2) the trial-level data is available so that we can estimate at least one reliability index. All these studies shared raw data publicly (Golubickis & Macrae, 2021; Navon & Makovski, 2021; Qian et al., 2020; Schäfer & Frings, 2019; Svensson et al., 2022) and did not deviate from the original experimental paradigm. Additionally, we identified five articles that did not have publicly available data but mentioned that data could be obtained upon request (Bukowski et al., 2021; Cheng & Tseng, 2019; Kolvoort et al., 2020; Martínez-Pérez et al., 2020; Xu et al., 2021). One of these articles indicated that data were shared on the Open Science Framework (OSF) platform (https://osf.io/pcv3u/), but

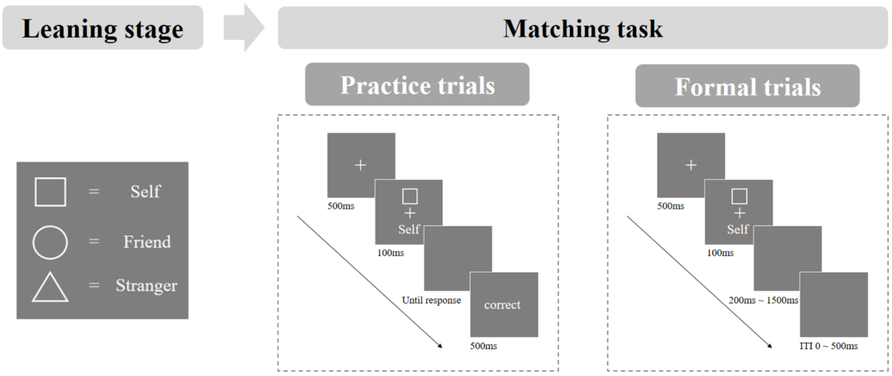

Fig. 1 Procedure of the original SPMT in the Experiment 1 Sui et al. (2012). *Note*: The relation between shape-label pairs is counter-balanced between participants.

the repository was found to be empty (Bukowski et al., 2021). We included datasets with raw data that were accessible to us. It is worth noting that the nature of the research culture discourages direct replications (Makel et al., 2012); thus, all datasets included in our analysis involved some degree of modification to the original design, such as incorporating additional independent variables or using different experimental materials (see our preregistration for details). 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 datasets used are described in Table 1.

3 Analysis

In our initial 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, during parameter recovery, we discovered that the parameters provided by the"hausekeep" package differed significantly from those obtained with the original HDDM package (Wiecki et al., 2013) used in previous studies (Golubickis et al., 2017). As a result, we made the decision to replace the original package with the "RWiener" package (Wabersich & Vandekerckhove, 2014), which we found to yield the most comparable results to the HDDM package. For detailed model comparison results, please refer to the supplementary materials. All the analyses in this paper are performed using the statistical software R (Dalgaard, 2010). The research flow of the current study is visually represented in Fig. 2.

Table 1 Datasets Information

		Independent Variable						SPE Indices						Reliability	
Paper	Exp.	IV 1	IV 2	IV 3	IV 4	Sample Size	# of Trials per Condition	RT	ACC	d	Eff	v	z	ICC	SHR
Hu et al. (2023)	1	Matching	Identity	Emotion Control, Neutral, Happy, Sad	Session	34	60	✓	✓	✓	✓	✓	✓	✓	✓
Constable and Knoblich (2020)	1	Matching	Identity	Switch Identity Partner, Stranger	Phase	92	40	✓	✓	✓	✓	✓	✓		✓
Constable et al. (2021)	2	Matching	Identity Self; Stranger	– –	– –	51	24	✓	✓	✓	✓	✓	✓		✓
Qian et al. (2020)	1	Matching	Identity Self; Stranger; Celebrity	Mood (Session)	– –	24	24	✓	✓	✓	✓	✓	✓		✓
	2	Matching	Identity Self; Celebrity	Cue With, Without	– –	25	50	✓	✓	✓	✓	✓	✓		✓
Schäfer and Frings (2019)	1	Matching	Identity Self, Mother, Acquaintance	– –	– –	103	24	✓	✓	✓	✓	✓	✓		✓
Golubickis and Macrae (2021)	1	Matching	Identity	Presentation Mixed; Blocked	– –	30	30	✓	✓	✓	✓	✓	✓		✓
	1	Matching	Identity	– –	– –	13	60	✓	✓	✓	✓	✓	✓		✓
Navon and Makovski (2021)	3	Matching	Identity Self; Father; Stranger	– –	– –	27	60	✓	✓	✓	✓	✓	✓		✓
	4	Matching	Identity	– –	– –	26	60	✓	✓	✓	✓	✓	✓		✓
	1	Matching	Identity Self; Friend	– –	– –	20	50	✓	✓	✓	✓	✓	✓		✓
Svensson et al. (2022)	2	Matching	Identity Self; Friend	Frequency Self > Friend	– –	24	100	✓	✓	✓	✓	✓	✓		✓
	3	Matching	Identity Self; Friend	Frequency Self < Friend	– –	25	100	✓	✓	✓	✓	✓	✓		✓
Cheng and Tseng (2019)	1	Matching	Identity	Go/No-go	– –	22	75	✓	✓	✓	✓	✓	✓		✓
	2	Matching	Identity	Go/No-go	– –	26	75	✓	✓	✓	✓	✓	✓		✓
	3	Matching	Identity	Go/No-go	– –	22	75	✓	✓	✓	✓	✓	✓		✓
Bukowski et al. (2021)	1	Matching	Identity	Imitation	– –	91	60	✓	✓	✓	✓	✓	✓		✓
	2	Matching	Identity	Imitation	– –	109	60	✓	✓	✓	✓	✓	✓		✓
Kolvoort et al. (2020)	1	Matching	Identity	Delay, 0, 40, 120, 700	– –	31	25	✓	✓	✓	✓	✓	✓		✓
Martínez-Pérez et al. (2020)	1	Matching	Identity	Simulation	– –	90	40	✓	✓	✓	✓	✓	✓		✓
Xu et al. (2021)	1	Matching	Identity	Feedback	Sex	105	60	✓	✓	✓	✓	✓	✓		✓
Woźniak et al. (2018)	1	Matching	Identity	Facial Gender Male; Female	– –	18	56	✓	✓	✓	✓	✓	✓		✓
	2	Matching	Identity	Facial Gender Male; Female	– –	18	60	✓	✓	✓	✓	✓	✓		✓
T. Liu et al. (2023)	1	Matching	Identity Self; Stranger	– –	– –	298	16	✓	✓	✓	✓	✓	✓		✓

Note: SPE: Self-Prioritization Effect, ICC: Intraclass Correlation Coefficient, SHR: Split-Half Reliability

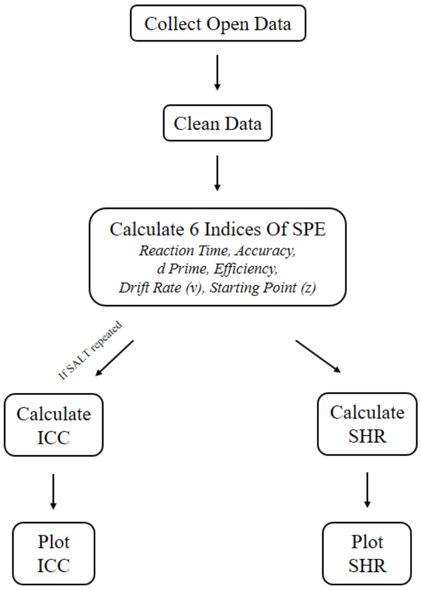

Fig. 2 Roadmap of the current study. *Note*: SPE: self-prioritization effect; d-prime is the sensitivity index under the Signal Detection Theory; drift rate (v) and starting point (z) are parameters derived from the Drift-diffusion Model; ICC: Intraclass Correlation Coefficient, SHR: Split-half Reliability.

3.1 Data Pre-processing

In total, we gathered 18 publicly available datasets, as mentioned earlier and presented in Table 1. We pre-processed the secondary data using the following criteria:

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. Behavioural Data Exclusion Criteria

   (i) Trials with no response or wrong key press is excluded from the analysis,

   (ii) The practice trials is excluded from the formal analysis,

   (iii) Participants with any of the conditions with zero accuracy is excluded from the analysis,

   (iv) The data under conditions other than the "control condition" would not be used in the current study.

3.1.1 Calculation of SPE

For each dataset, we calculated six indices for each experimental condition: Mean RT (MRT), accuracy (ACC), d-prime (d'), efficiency (η), drift rate (v), and starting point (z). Mean RT and ACC are obtained directly from the datasets, while d' and η are calculated based on Mean RT and ACC using a simple formula (see Table 2).

3.1.2 Estimating the Reliability

Split-half reliability. We calculated the split-half reliability of the six indices using four types of split-half reliability measures: odd-even, front-back, permutation, and Monte Carlo (Kahveci et al., 2022; Pronk et al., 2022). The odd-even split divides trials into odd and even numbered sequences, while the front-back split divides the first and second halves of trials. The permutation split shuffles the trial order and randomly assigns each half to a group. The Monte Carlo split-half is similar to the permutation split-half, but it repeats the process thousands of times to calculate the average and 95% confidence interval of the split-half reliability. This study will primarily use Monte Carlo split-half to determine the split-half reliability of SPMT for its robustness (Pronk et al., 2022). The results of the other three split-half methods is presented in the supplementary materials.

First, the data is stratified according to Session (if applicable), Matching, and Identity. If the data is not stratified, directly splitting it in half will result in an uneven distribution of trials for each experimental condition in the two halves, which can lead to an overestimation or underestimation of split-half reliability. Therefore, once the data is stratified, we split it into two halves. For example, when using Monte Carlo Split-Half, we randomly split the data into two halves. Then we repeat this process 1000 times. This will result in 1000 pairs of two halves of the data. Next, we use these 1000 pairs of data to calculate 1000 Pearson correlation coefficients, and then obtain the average and 95% confidence interval of the Monte Carlo split reliability. First-second split, odd-even split, and permutated split are similar to Monte Carlo method, but they only perform one split, so only one split-half reliability is obtained without an interval estimate of the split-half reliability.

Test-Retest Reliability (ICC). We assessed the test-retest reliability of the six indices in our dataset that involved multiple experiment sessions by calculating the Intraclass Correlation Coefficient (ICC). To perform this analysis, we utilized the "psych" package as described by (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 takes into account both the correlation and agreement between multiple measurements, making it a more comprehensive measure of test-retest reliability. Within the ICC family, we specifically employed 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; Liljequist et al., 2019). For the calculation of ICC2 estimates, the formula is:

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

Table 2 Indices in SPMT and corresponding SPE calculation

Indices	Indices Calculation	SPE Calculation Based on Indices	Source
Mean RT (MRT)	Total RT/Total Responses	$\text{RT}_{\text{other-matching}} - \text{RT}_{\text{self-matching}}$	Sui et al. (2012)
Accuracy (ACC)	# of Correct Responses/Total Responses	$\text{ACC}_{\text{self-matching}} - \text{ACC}_{\text{other-matching}}$	Sui et al. (2012)
d-prime (d')	$\mathcal{Z}$ (Hits) $-$ $\mathcal{Z}$ (False Alarms)	$d'_{\text{self-matching}} - d'_{\text{other-matching}}$	Sui et al. (2012)
Efficiency (η)	MRT/ACC	$\eta_{\text{self-matching}} - \eta_{\text{other-matching}}$	Humphreys and Sui (2015) and Stoeber and Eysenck (2008)
Drift Rate (v)	Parameters decomposed from RT based on DDM	$v_{\text{self-matching}} - v_{\text{other-matching}}$	Golubickis et al. (2017)
Starting Point (z)		$z_{\text{self-matching}} - z_{\text{other-matching}}$	Golubickis et al. (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.

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:

$$\text{ICC2k} = \frac{MSBS - MSE}{MSBS + \left(\frac{MSBM - MSE}{n}\right)}. \tag{2}$$

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

5 Discussion

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. However, despite its importance, this practice is not yet widely adopted (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 indices related to the self-prioritization effect (SPE) in the self-perceptual matching task (SPMT). We re-analyzed data from 18 datasets across 11 articles by employing the intraclass correlation coefficient (ICC2,ICC2k) and split-half reliability for this purpose. Our analysis of these datasets collectively demonstrate that RT yield better results compared with other indices, and the result varies between different associations. The test-retest reliability results suggest that Response time (RT) and efficiency score consistently exhibits high ICC2k and low ICC2 across datasets (xxx). However, measurements based on accuracy and DDM yield varying outcomes. Overall, the indices related to the SPE in the SPMT are more suitable for group-level analysis rather than assessing individual-level variation.

In terms of split-half reliability, the results indicate variations between the target and indices. Specifically, RT yields superior results compared to other indices, suggesting it is the most reliable indicator for accurately assessing and distinguishing between the self and other targets in the SPMT. Furthermore, when examining the split-half reliability for the self-other difference, self-stranger is the highest among other comparisons. Such finding suggests that the measurement of this particular difference remains consistent across participants and studies, indicating the systematic processing difference. The self-friend differences demonstrate the lowest reliability, indicating that the measurements obtained from the indices are not stable or consistent when split into two halves. This finding suggests that distinguishing between the self and friend in the paradigm is challenging. The aforementioned result aligns with previous studies conducted using SPMT, consistently demonstrating the establishment of a reliable self-advantage. This advantage is observed when the shapes are associated with the self, in comparison to when they are linked to an unfamiliar person or a neutral label(Feng et al., 2020; Sui et al., 2012).

Although RT yield the highest split-half reliability among other indices, the result is still below a commonly considered excellent reliability (). The low test-retest reliability may suggest that the indices in SPMT is subject to random error and inconsistent performance. Several plausible reasons can be identified. Firstly, a potential contributor to low split-half reliability could be the insufficient number of trials per condition. A recent study by Kucina et al. (2023) has highlighted the significance of the number of trials in cognitive tasks for determining reliability. The findings of the study revealed that increasing the number of trials resulted in greater magnitudes of conflict effects and individual differences. Consequently, this led to improved reliability when compared to previous archival data. Specifically, in the case of gamified Flanker task, the study identified that achieving satisfactory reliability required 48 or fewer trials, while achieving a higher level of reliability necessitated 72 trials. Therefore, incorporating a higher number of trials in future employment of the SPMT paradigm may enhance the split-half reliability by enhancing the consistency of measurement. Second, 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). SeriAal 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.

The discrepancy between the high ICC2k and low ICC2 suggests that thew SPMT is more influenced by between-participant variability than within-participant variability (Hedge et al., 2018; Liljequist et al., 2019). It is common for behavioral paradigm to have such result pattern, as demonstrated in previous research testing other cognitive paradigms such as Flanker, Simon, or Stroop (Clark et al., 2022; Mollon et al., 2017). There are various reasons for this pattern. First, one significant factor could be the prevalence of practice effect, particularly if the practice effect is large enough to cause a substantial change in participants' performance between measurement occasions, it can introduce additional variability in the measurements. This increased variability may lower the ICC2 (Oswald et al., 2015; Siegelman et al., 2017). The presence of a practice effect underscores the need for alternative measures that can consistently capture performance nuances and reveal individual differences more sensitively (Hedge et al., 2018). To address this limitation, researchers can consider incorporating additional performance metrics, such as composite RT-accuracy scores.

By including RT alongside accuracy, a more stable assessment of participants' abilities can be achieved, allowing for greater ICC2. Second, behavioral paradigms are susceptible to factors such as external conditions, contextual differences etc.., which contribute to greater within-participant variability and lower ICC2 values. However, when averaging performance between different individuals, the task could still exhibit good consistency, resulting in higher ICC2k values. It's important to note that low ICC values should not be solely interpreted as a measure of a test's overall quality but rather as an indication of the types of questions it can effectively address. In practical terms, the results suggest that the SPMT is better suited for distinguishing performance differences between individuals or groups, rather than capturing consistent performance within the same individuals over time. Thus, the SPMT may be particularly useful for studying inter-individual variability or conducting group-level comparisons, rather than tracking individual-level changes or stability. Therefore, we recommend that researchers take these factors into consideration when investigating individual differences in performance using the SPMT.

Our study has a few limitations that should be acknowledged. Firstly, although we made efforts to enhance sample diversity by including open data as much as possible, it is important to note that a majority of our samples still consisted of individuals from what is commonly referred to as"wired" populations (Rad et al., 2018). Therefore, our findings may not be fully representative of the broader population, and a more diverse sample is needed to ensure greater generalizability. Additionally, it is important to highlight that the majority of the studies included in our analysis focused on adults from healthy populations. Hence, further investigation is needed to determine the reliability of the SPMT across different age groups and clinical populations.Secondly, 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 indices. Consequently, it is recommended that future research focuses 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, which 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, the current study find that RT-base measurements proved more robust than accuracy ones. Moreover, SPMT is more suitable for group-level analysis rather than assessing individual-level variation. The findings of our study offer significant insights into the reliability of SPMT, shedding light on important factors that require careful consideration when interpreting the reliabilities. These findings also have 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.

Acknowledgments

The present research is support by xxx.

Author Contributions

HCP contributed to the conception and supervision of the study. JS contributed to fund raising, HCP contributed to data collection. ZL, ZYR and HMZ performed the data pre-processing, analysis and visualize the results. In addition, ZL, JS, HMZ and HCP contributed to the discussion of the results and the drafting of the final manuscript. All authors critically revised 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 (Dataset 0) 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

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., 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

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). 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

Dalgaard, P. (2010). R development core team (2010): R: A language and environment for statistical computing.

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. 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

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

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

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., 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.

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., Peng, K., & Sui, J. (2023). Data for training effect of self prioritization[ds/ol]. v1. *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

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

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

Ivaz, L., Costa, A., & Duñabeitia, J. A. (2016). The emotional impact of being myself: Emotions and foreign-language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(3), 489. https://doi.org/10.1037/xlm0000179

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.

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., 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

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

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

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. https://doi.org/10.1177/2515245919879695

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

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

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.

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

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

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). 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). The self-reference effect in memory: A meta-analysis. *Psychological bulletin*, *121*(3), 371–94. 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

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

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

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