**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:***

Recent years have witnessed a growing focus on the reliability of cognitive tasks, driven in part by the reliability paradox. This paradox stems from the observation that while cognitive tasks yield consistent experimental effects, they do not exhibit the same reliability when assessing individual differences. Here we investigate the reliability of the Self Perceptual Matching Task (SPMT), a widely used tool for investigating the cognitive processes underlying the self-prioritization effect (SPE), a effect that people perform better when stimuli are associated to the self than when they are to others. In this preregistered study, we evaluated the reliability of 24 SPE measures from 17 datasets (N = 805), all utilizing the SPMT. We calculated Monte-Carlo based split-half reliability (r) and intraclass correlation coefficient (ICC2) for each SPE measure. Our findings revealed a robust group-level SPE effect across datasets. However, when it comes to individual differences, SPE measures derived from reaction times (RT) and Efficiency exhibited relatively higher, compared to other SPE measures, but still unsatisfied split-half reliability (approximately 0.6). Similarly, for the reliability across multiple time points, as assessed by ICC2, RT and Efficiency demonstrated low levels of test-retest reliability (close to 0.5). These outcomes uncover the presence of a reliability paradox in the context of SPMT-based SPE assessments. While nearly all the measures of SPE displayed robust experimental effects, their reliability are low as a measurement of individual differences. We discussed the implications of the current study for future studies.

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

# **1 Introduction**

The self-prioritization effect (SPE) reflects individuals' biased responses towards self-related information in comparison to that related to others. This effect has been found in a broad range of cognitive tasks. This phenomenon, documented in 1950s (Cherry, 1953), holding a central position within cognitive psychology and underscoring a core facet of human cognition and self-awareness (Sui & Humphreys, 2017). Despite SPE is often argued to be a self-specific effect, it has been challenging to be disassociated from familiarity effect. That is, the self-related stimuli, such as own objects, own faces (Keenan et al., 2000; Kircher et al., 2000; Turk et al., 2002), own voices (Hughes & Harrison, 2013; Payne et al., 2021), or own names (Constable, Rajsic, et al., 2019) are usually more familiar to participants than those other-related stimuli. To overcome such limitation, Sui et al. (2012) introduced the Self-Perceptual Matching Task (SPMT), where the self-relatedness (and other-relatedness) is acquired in the lab. In this task, participants first associate geometric shapes with person labels (e.g., circle = you, triangle = best friend, square = stranger) and then perform a matching task, judging whether a shape-label pair presented on the screen match the acquired relationship. A typical pattern from this task is that shapes associated to the self enjoy an advantage over shapes related to others. This SPE from SPMT was subsequently observed by many researchers (Constable, Elekes, et al., 2019; Golubickis et al., 2020; Golubickis et al., 2017; Hu et al., 2020), highlighting the robustness of the effect.

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

In recent years, driven by a growing interest in employing cognitive tasks to assess individual differences, researchers have turned their attention to evaluating the reliability of cognitive tasks (e.g., Karvelis et al., 2023). However, existing findings have raised concerns about the reliability of many cognitive tasks (Hedge, Powell, & Sumner, 2018). A considerable body of research has highlighted the moderate to low level reliability found in the cognitive task measurements (Clark et al., 2022; Enkavi et al., 2019; Green et al., 2016). For instance, Hedge et al. (2018) report a range of test-retest reliabilities pertaining to frequently employed experimental task metrics (such as Stroop and Stop-Signal Task), with a notable prevalence of discrepancy between the low reliability for individual differences and the robustness of the experimental effects. This discrepancy, named as "reliability paradox" (Logie et al., 1996), had gain much attention in recent years (Hedge, Powell, Bompas, et al., 2018; Hedge, Powell, & Sumner, 2018). As other cognitive tasks, SPMT is also employed by researchers as a measure of individual diffrences in SPE. For example, a recent study examine the individual difference of SPE and how these individual differences are correlated to brain network (Zhang et al., 2023). Likewise, in clinical investigation, the SPMT has been incorporated to assess deviations in self-processing among specific populations, including individuals affected by autism or depression (Hobbs et al., 2023; Liu et al., 2022). This trend calls for assessing the reliability of SPMT as a measurement of SPE.

Further, the varibilities in quantifying SPE call for a comphensive examination of the relability of different SPE measures. As simple as the SPMT, there are multiple approaches to quantify the SPE, encompassing various indicators and baselines. In a typical SPMT experiment, two direct outcomes are generated: reaction times (RTs) and choices. The mean RTs and accuracy are two most widely used indicators of SPE. Several indicators can be derived from reaction times and choice: 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). In addition to the variability of indicators, SPE can be estimated by calculating the difference between self condition with different baselines. Indeed, the selection of baselines varies across studies, including “Close other” (e.g., Friend) (Navon & Makovski, 2021; Svensson et al., 2022)), “Stranger” (Constable et al., 2021; Orellana-Corrales et al.), “Celebrity” (e.g., “LuXun”) (Qian et al., 2020), and “Non-person” (e.g., None)) (Schäfer & Frings, 2019). As a result, two pivotal questions regarding the reliability of the SPMT remain unresolved: (1) given the variability of indicators (RT, ACC, *d’, η, v, z*) and choice of baseline conditions ("Close other”, “Stranger”, “Celebrity”, and “Non-person”), which way of quantifying SPE is the most reliable one(s)? (2) Is the SPE measured by SPMT suitable for assessing individual differences? Addressing these questions is crucial for establishing the formal reliability analysis of SPMT measurements, allowing for accurate assessment of the SPE and its implications in various domains.

To address these two questions, the present study investigated the reliability of SPE measures computed using different indicators under various baseline conditions in the SPMT. This was achieved by re-analyzed 17 independent datasets (N = 805) from 9 papers and 2 unpublished projects that employees SPMT. In order to comprehensively assess the SPE measures derived from SPMT, we examined six indicators (RT, ACC, *d’*, *η*, *v*, *z*) under four baseline conditions (“Close other”, “Stranger”, “Celebrity”, and “Non-person”). We first assessed the experimental effect of theses SPE measures using meta-analytical assessments. Given the methods available for evaluating the reliability of cognitive tasks, we examine the individual level consistency via permutation-based Split-half reliability (*r*), and consistency of task performance over time using Intraclass correlation coefficient (ICC2, Two-way random effect model). The findings of our study provide valuable insights into the reliability of SPMT and its indicators, having the potential to facilitate the future utilization of SPMT in research, clinical settings, and personal performance monitoring.

# **2 Methods**

## 2.1 Ethics information

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

## 2.2 Experimental Design

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.

Here we provided a detailed overview of the original experimental design of SPMT, as described in the Experiment 1 by Sui et al. (2012). The original SPMT consisted of two stages (refer to Fig. 1). In the first instruction stage (learning stage), participants were instructed to associate three geometric shapes (circle, triangle and square) with three labels (self, friend, and stranger) for approximately 60 seconds. The shape-label associations were 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.

A diagram of a task

Description automatically generated

**Fig. 1.** Procedure of the original SPMT in the Experiment 1 (Sui et al., 2012).

## 2.3 Datasets Acquisition

Initially, two datasets that employed the SPMT were available to us: one from an unpublished project conducted in our laboratory (Hu et al., 2023), for which we provide a detailed experiemntal design in the supplemetray, and the other provided by our collaborators (Liu et al., 2023). Concurrently, we are conducting a meta-analysis on SPE using the SPMT (pre-registration available at <https://osf.io/euqmf>). During this process, we identified an additional thirteen papers with datasets potentially suitable for our present study. The selection of these papers was based on specific criteria:

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

Among the thirteen papers included, seven papers made their trial-level 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 six 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 trial level data (Kolvoort et al., 2020; Woźniak et al., 2018; Xu et al., 2021). However, in one article, the author did not provide the explanation of the shape and label in the original data (Kolvoort et al., 2020). As a result, we are unable to analyze the raw data in this context. Two papers provided us only with descriptive results (Cheng & Tseng, 2019; Martínez-Pérez et al., 2020), which unfortunately could not be used for calculating reliability. Additionally, one paper referred to data being shared on the Open Science Framework (OSF) platform (<https://osf.io/pcv3u/>) (Bukowski et al., 2021), but we found that the repository was empty, making it ineligible for the current analysis.

In total, our analysis comprised raw data from nine papers and two unpublished projects from our laboratory and collaborators. It is important to highlight that the research culture discourages direct replications (Makel et al., 2012). As a result, all the datasets included in our analysis underwent some degree of modification to the original design (e.g., change shapes, modify sequence) as well as including the incorporation of additional independent variables (refer to Table 1 for specification). For our analysis, we focused exclusively on using data that adhered to the original design of SPMT, without incorporating any stimuli that could potentially trigger a familiarity effect. Additionally, we specifically utilized data from control conditions in the studies, excluding any conditions that involved the presence of other independent variables (e.g., mood changes). In the end, we were able to incorporate 17 independent datasets from the above-mentioned papers and projects. Nonetheless, not all studies incorporated retest sessions. If a publicly available datasets did not include retest session with SPMT, we excluded it from calculating the Intraclass Correlation Coefficient and only considered the split-half reliability. The details of the included studies and conditions in the datasets are described in Table 1.

Table 1. Dataset Information

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

*Note*. Study represents different studies from a single article; IV: independent variable. For IV3 and IV4, we only included the baseline conditions that are similar to the original design in Sui et al. (2012), which were highlighted in **BOLD font**. If other variables that could be counterbalanced are indicated by underscores, we will solely utilize these variables as stratification variables during the split-half process

## 2.4 Analysis

Analysis plans for this study were preregistered on OSF (https://osf.io/zv628). All analyses in this paper were performed using the statistical software R (R Core Team, 2023). The drift rate (*v*) and starting point (*z*) of the drift-diffusion model (DDM) was obtained using the “RWiener” package (Wabersich & Vandekerckhove, 2014).

The visual representation of the current study's roadmap can be found in Fig. 2 and will be further elucidated in the subsequent sections. After collecting the data from each independent study, we performed data cleaning and then calculated the six indicators as well as the twenty-four SPE measures computed using different indicators and baseline conditions. Then, we computed the effect sizes (Hedges’ *g*) of these SPE measures. In addition to effect sizes, we calculated the split-half reliability of these SPE measures. If there were test-retest data, we also calculated the test-retest reliability using the intraclass correlation coefficient (ICC2, see Estimating the Reliability Section for details).



**Fig. 2** Roadmap of the Current Study.

*Note:* \*Only one paper have Celebrity and Nonpersons baseline, thus no included in the meta-analysis

## 2.4.1 Data Pre-processing

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

1. Participant exclusion criteria
2. Participants who had wrong trial numbers due to 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 where the keypress occurs outside the two required keys and non-responsive trials are excluded from the analysis,
7. the practice trials are excluded,
8. the experimental design involved independent variables more than self-referential and matching, e.g., included valence of emotion as a third independent variable.

## 2.4.2 Calculating the Indicators and SPE Measures

For each study, we calculated six indicators for each experimental condition: Mean Reaction Time (RT), Accuracy (ACC), Sensitivity Score (*d′*), Efficiency (*η*), Drift Rate (*v*), and Starting Point (z). Reaction Time and Accuracy are obtained directly from the datasets, while Sensitivity Score and Efficiency are calculated based on Reaction Time and Accuracy. SPE Measures were computed using different indicators under four baseline conditions (see Table 2).



## Estimating the Robustness of SPE

**Calculation of Effect Sizes.** In this study, the robustness of experimental effects (group-level effect) of SPE in SPMT is calculated using meta-analytical assessments. In our analysis of effect sizes, we employed a random effects model as the primary approach, given the anticipated heterogeneity among participant sample (Page et al., 2021). The effect size index used for all outcome measures is Hedges’ *g*, a correction of Cohen's *d* that accounts for bias in small sample sizes (Hedges & Olkin, 2014), which represents the magnitude of the difference between the self and baseline condition. Descriptive statistics, including sample size, mean, and standard deviation, are employed to calculate Hedges’ *g* from the datasets.

When calculating Hedges' *g*, we have converted the effect size vlalue of Reaction Time and Efficiency into absolute values. Conversely, for all indicators, a positive effect siz indicates a bias towards associating stimuli with the self rather than with baseline associations. For the estimation and interpretation of effect sizes, effect size around 0.2 is interpreted as small effect size, around 0.5 as medium effect size, and around 0.8 as large effect size (Fritz et al., 2012; Hedges & Olkin, 1985).

## Estimating the Reliability of SPE

**Split-half Reliability.** We assessed the split-half reliability by first splitting the trial-level data into two halves and calculating the Pearson correlation coefficients. To ensure methodological rigor, we used four data splitting approaches for splitting the trial-level data: first-second, odd-even, permutated, and Monte Carlo (Kahveci et al., 2022; Pronk et al., 2022). The first-second approach splits trials into the first half and the second half. The odd-even approach splits the trials into sequences based on their odd or even numbers. The permutation approach shuffled the trial order and randomly assigned trials to two halves. The Monte Carlo approach is similar to the permutation approach, but iterates the process multiple times, usually thousands of times, to calculate the average and 95% confidence intervals of the split-half reliability.

In our analyses, we first stratified the trial-level data for each participant in the study based on experimental conditions. For example, in the case of a 2 by 3 within-subject design, we stratified the data based on the two independent variables: identity (self, stranger, friend) and matching (match, non-match). Subsequently, we applied the four splitting approaches (Pronk et al., 2022). When using Monte Carlo approach, 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 for these 5000 pairs. After that, we calculated the mean and 95% confidence intervals of the 5000 correlations coefficients. The first-second split, odd-even split, and permutated split are similar to the Monte Carlo approach except each of these approaches only results in a single reliability coefficient. Finally, after computing the split-half reliability coefficients for each dataset, substantial variations were observed across the datasets. To derive a more accurate estimation of the average split-half reliability for each SPE measures, we synthesized these reliability coefficients via a meta-analytical approach. We weighed the reliability coefficients based on the trial numbers of each study since the number of trials typically significantly influences the reliability of cognitive experiments (Kucina et al.2023) (see also Supplementary Fig. 4 for our exploratory analysis). The weighted-average reliabilities was calculated use the “aggregate.escalc” function in the “metafor” Package (Viechtbauer, 2010). We report the synthesized split-half reliability and its 95% confidence interval in the results.

Although there is no strict criterion for defining the level of split-half reliability for psychological and educational measures, a widely accepted guideline for split-half reliability coefficient is that a value of 0.70 is “acceptable”, and a value greater than 0.8 means excellent reliability (Cicchetti & Sparrow, 1981; Kupper & Hafner, 1989).

**Test-Retest Reliability (ICC).** The Intraclass Correlation Coefficient (ICC) serves as a widely recognized measure for evaluating test-retest reliability (Fisher, 1992). Differing from the Pearson correlation coefficient, which primarily quantifies the linear association between two continuous variables, the ICC extends its prowess to scenarios involving multiple measurements taken on the same subjects, while also considers both the correlation and agreement between multiple measurements, making it a more comprehensive measure of test-retest reliability (Koo & Li, 2016).

Since our primary aim is to evaluate the appropriateness of the SPMT in assessing individual differences and repeated administration, to achieve this objective, we assessed the test-retest reliability of the six indicators for our dataset that involved test-retest sessions using the function “ICC” in the “psych” package (Revelle, 2017). We focused on using the Two-way random effect model (ICC2) within the ICC family (Chen et al., 2018; Xu et al., 2023). ICC2 gives an estimate of the proportion of total variance in measurements that is attributed to between-subjects variability (individual differences) and within-subjects variability (variability due to repeated measurements) (Xu et al., 2023). For the calculation of ICC2 estimates, the formula is:

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.

The traditional benchmarks for interpreting ICC values are as follows: ICC less than 0.50 suggests poor reliability; ICC between 0.50 and 0.75 suggests moderate reliability; ICC between 0.75 and 0.9 suggests good reliability; ICC above 0.9 suggests excellent reliability (Cicchetti & Sparrow, 1981; Kupper & Hafner, 1989).

# **3 Deviation from Preregistration**

We adhere to our pre-registration plan as much as possible, however, there are a few differences between the current report and pre-registration document. First, in our initial preregistration plan, we did not anticipate conducting an analysis on the group-level effect of SPE due to the perceived robustness of the effect across a diverse range of research. However, as our study progressed, we recognized the value in providing a more comprehensive assessment. Thus, we include a estimation of pooled effect sizes across included study to represent the group-level effect. Second, we used a different algorithm for estimating parameters of the drift-diffusion model. In the registration, we planned to estimate the drift rate (*v*) and starting point (*z*) of the drift-diffusion model using the “fit\_ezddm” function from the “hausekeep” package (Lin et al., 2020). This function served as a wrapper for the EZ-DDM function (Wagenmakers et al., 2007). However, we found that this algorithm is flawed when estimating parameter *z* during in parameters recovery (details provided in the Supplementary Materials). After comparing the 5 algorithms, we found that the "RWiener" package (Wabersich & Vandekerckhove, 2014) achieved a favorable balance between accuracy, confidence interval width, and computational efficiency, making it the most suitable choice for our analysis. Nevertheless, for transparency, we have included the results from ezDDM in the supplementary materials (see Supplementary, Fig. 2-4). Third, we did not explicitly state in the preregistration report that we would perform a weighted average of the Monte Carlo split-half reliabilities for all obtained studies. However, considering that the number of trials has a significant impact on reliability (Kucina et al., 2023), during the formal analysis, we assigned different weights to each study based on the number of trials and performed a weighted average of the split-half reliabilities. Forth, in our original preregistration, we outlined our intention to include both ICC2 and ICC2k in our data analysis. However, as our understanding of Intraclass Correlation Coefficients (ICC) improved, we realized that ICC2 is the appropriate index for our research purpose. More specifically, ICC2k was mentioned in the preregistration as an index of robustness of group-level effect, but it turned out to be another index of reliaiblity for individual differences. We corrected this mis-interpretation of ICC2k in the final report. Fifth, we conducted exploratory analysis using the data we collected to investigate the relationship between the number of trials, Monte Carlo split-half reliability, and effect size (Hedges’ *g*) (refer to Supplementary Fig. 5-7). Finally, the writing of the current manuscript was improved based on the preregistration. For example, in our preregistration, we include different baseline conditions when calculating SPE in the method section but did not mention this in our introduction and abstract. In this final report, we improved the writing and adjusted the introduction and abstract accordingly.

# **4** **Results**

In seventeen independent datasets, 14 of them contain data for the Close other baseline, 13 of them contain data for Stranger baseline, 1 of them have the data for Celebrities baseline, 1 of them has the data for Nonperson baseline. Since there is only one paper for 'Celebrity' and one for 'Nonperson,' their results is presented in the supplementary materials.

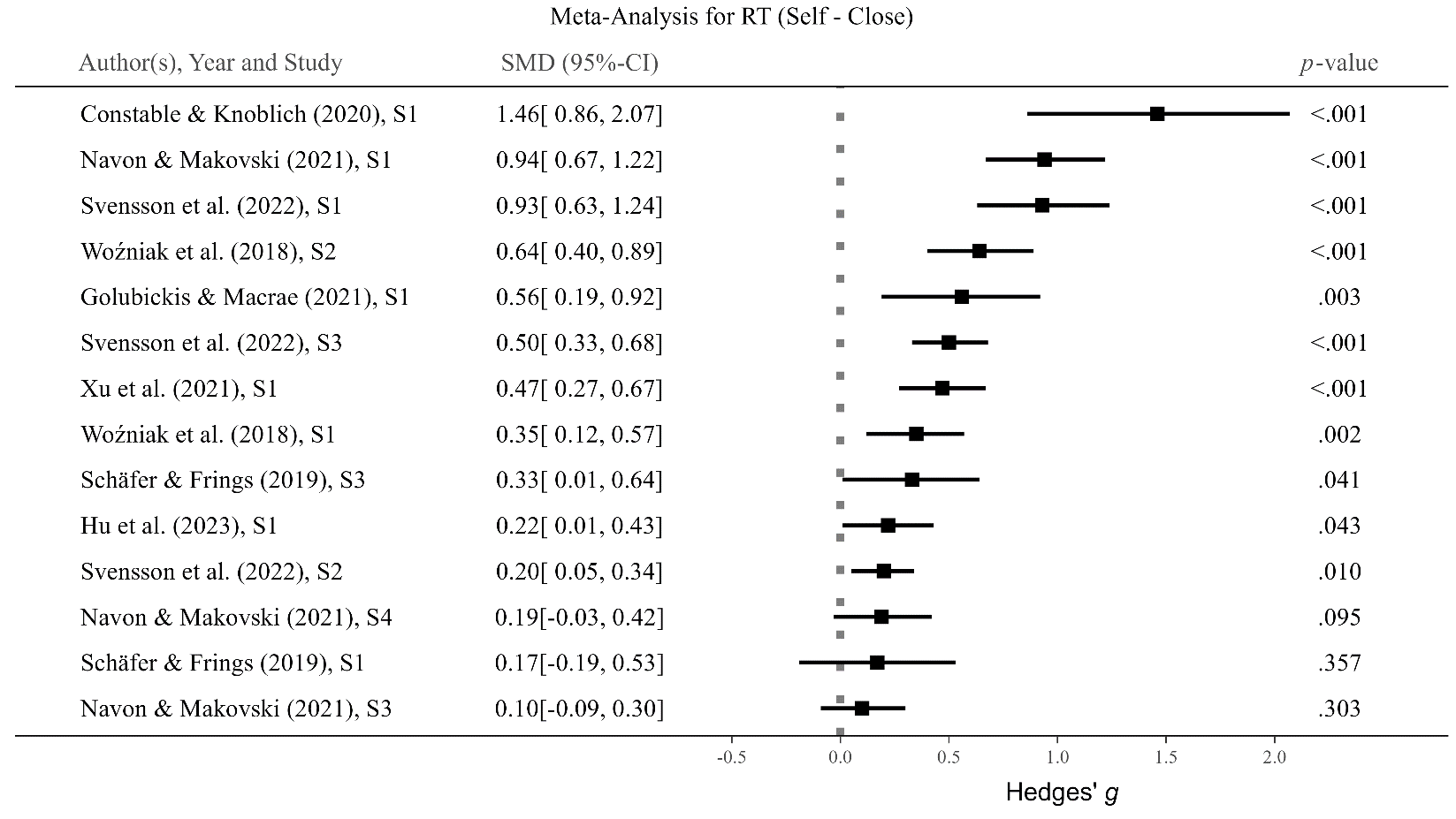
## 4.1 Group Level Effect of SPE

We conducted a meta-analytical assessment to examine the robustness of SPE as measured by SPMT. Because only one of the papers contains data for the "Self vs Celebrities" comparison, and one paper has data for the "Self vs none" condition, we exclude the Celebrities and none conditions from the analysis of overall effect sizes. The results, including Hedges’ *g* values and their 95% confidence intervals, along with the outcomes of the random effect model applied to SPE measures, are presented in Table 3. Furthermore, we have included the forest plot for RT within the main text, specifically in Fig. 3. For additional forest plots pertaining to other SPE measures, please refer to the supplementary material.

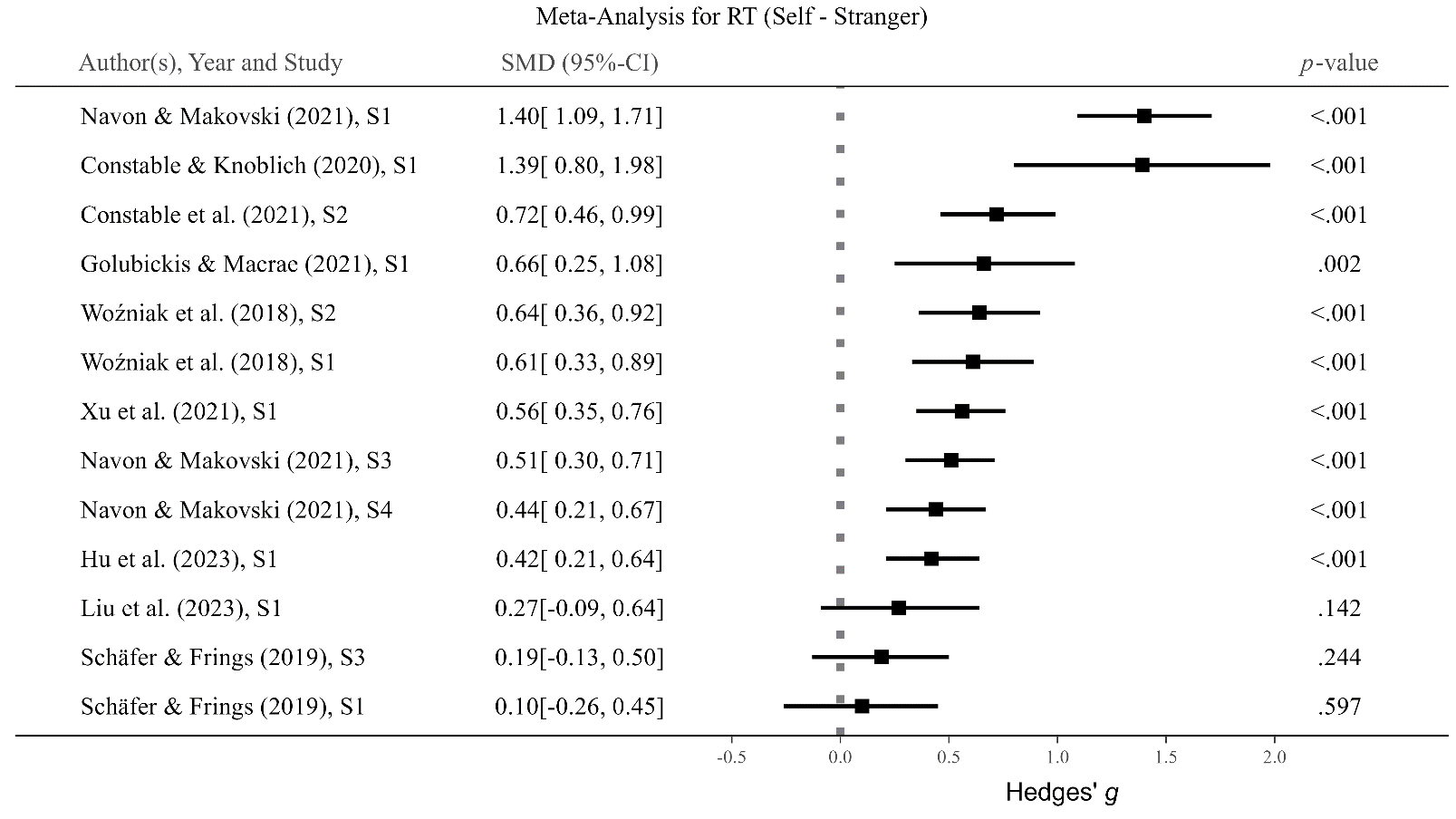
With the exception of the parameter *z*, all other parameters exhibit moderate to large effect sizes. These findings indicate a robust and substantial group-level effect of SPE in SPMT. There were strong evidence of homogeneity among studies. The value, all being greater than 75%, indicates high heterogeneity among studies, justifying the selection of the random effect model (Borenstein et al., 2021).

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Baseline** | **Indicators** | ***g* [95% CI]** | **No. of**  **Studies** | **Q** | ***p*** | ***I2*** |
| Close |  |  |  |  |  |  |
|  | RT | 0.47 [0.30, 0.63] | 14 | 68.67 | <.001 | 84.94% |
|  | ACC | 0.73 [0.42, 1.03] | 14 | 144.57 | <.001 | 92.87% |
|  | *d* | 0.44 [0.28, 0.59] | 14 | 81.96 | <.001 | 83.02% |
|  | *η* | 0.88 [0.50, 1.25] | 14 | 128.47 | <.001 | 94.67% |
|  | *v* | 0.54 [0.32, 0.76] | 14 | 142.79 | <.001 | 91.16% |
|  | *z* | 0.15 [-0.03, 0.33] | 14 | 122.30 | .11 | 88.95% |
| Stranger |  |  |  |  |  |  |
|  | RT | 0.59 [0.40, 0.78] | 13 | 55.30 | <.001 | 83.20% |
|  | ACC | 0.78 [0.48, 1.08] | 13 | 77.78 | <.001 | 88.60% |
|  | *d* | 0.35 [0.21, 0.50] | 13 | 47.81 | <.001 | 75.38% |
|  | *η* | 0.92 [0.56, 1.29] | 13 | 98.79 | <.001 | 93.30% |
|  | *v* | 0.44 [0.28, 0.59] | 13 | 50.98 | <.001 | 79.33% |
|  | *z* | 0.08 [-0.09, 0.24] | 13 | 70.48 | .37 | 84.44% |



**Figure 3a. Forest Plot of RT (Self – Close)**



**Figure 3b. Forest Plot of RT (Self – Stranger)**

## 4.2 Split-half Reliability

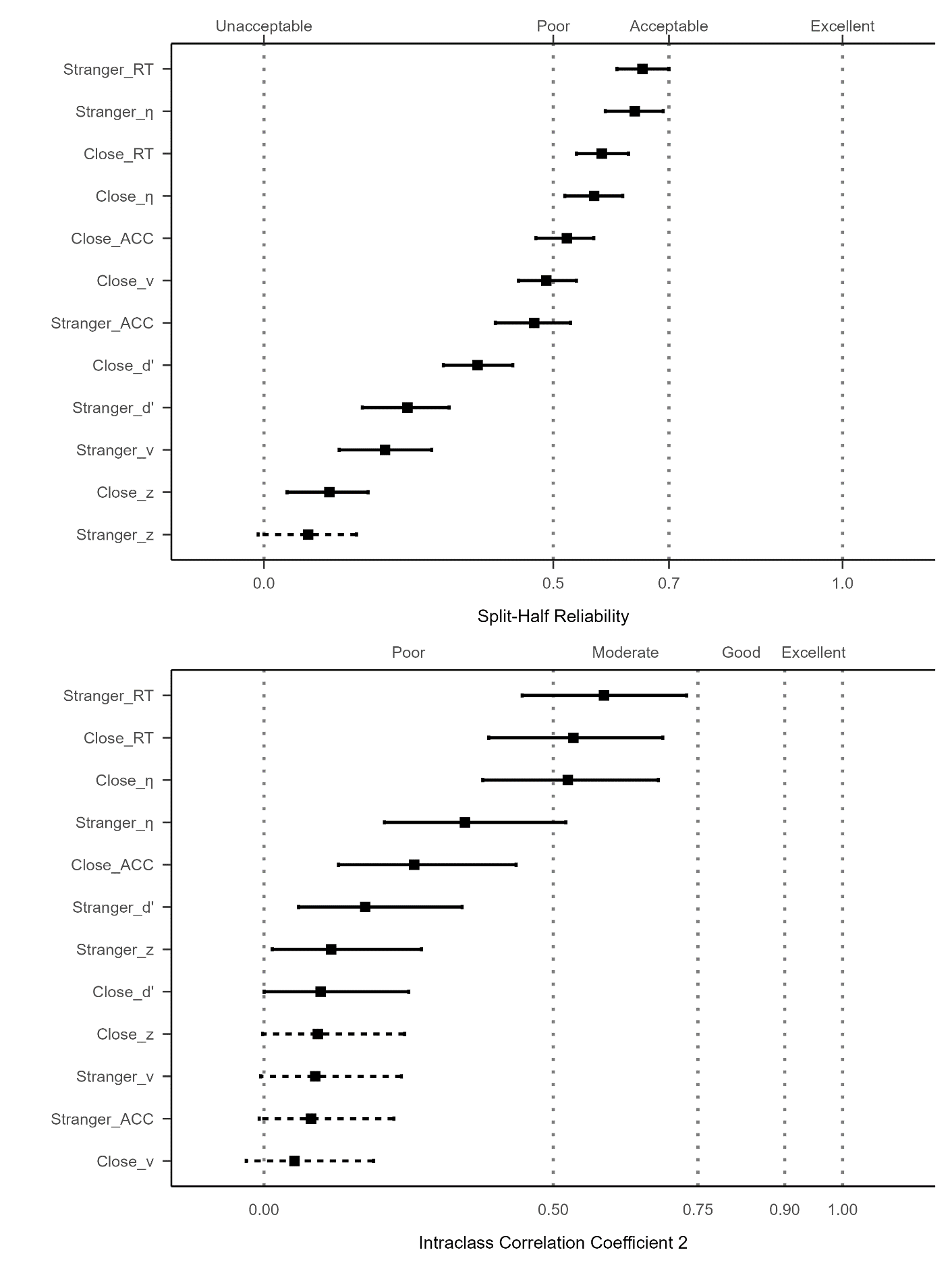
As described in method part, we applied four different approaches to split the data when calculating split-half reliability, namely the first-second, odd-even, permuted, and Monte Carlo methods. We presented the results from Monte Carlo split-half method in the main text due to its robustness (Pronk et al. (2022). The results of the other three split-half methods can be found in the supplementary materials (see Supplementary Fig. 2), except the odd-even, the other two also shared highly similar result pattern with the Monte Carlo methods.

The split-half reliabilities using Monte-Carlo method for different SPE measures is presented in Fig. 4 in the upper facet. The split-half reliabilities were weighed based on the trial numbers of each study. Among the SPE measures, the four highest ranking split-half reliabilities are as follows: Reaction Time (RT) with "Stranger" contrast (*r* = .65, SE = .02, *p* <.001, 95% CI [.61, .70]); Efficiency (*η*) with "Stranger" contrast (*r* = .64, SE = .03, *p* <.001, 95% CI [.59, .69]); RT with "Close other" contrast (*r* = .58, SE = .02, *p* <.001, 95% CI [.54, .63]); *η* with "Self vs Close" contrast (*r* = .57, SE = .02, *p* <.001, 95% CI [.52, . 62]). These SPE measures achieved a split-half reliability around 0.6 or higher, which is considered an acceptable level of reliability. For the remaining SPE measures, the reliability is around 0.5 or lower, indicating poor test-retest reliability. These include Accuracy (ACC), Sensitivity Score (*d’*), Drift Rate (*v*), and Starting Point (*z*) under four baselines. It is notice that almost all the split-half reliability of *z* using different baselines are around 0, which suggests a complete lack of reliability.

## 4.3 Test-retest Reliability

It is important to note that we could only calculate ICC for the study in our labortory (Hu et al., 2023) since all other datasets did not include re-test sessions. For this dataset, we only have 2 baslines, the “Close other” and “Stranger”. To test the robustness of the results, we also explored one additional datasets that included re-test session but devivated quite strongly from the original SPMT (see Supplementary Fig. 4).

The ICC2, which measure the reliability for individual differences, aligns with the findings observed in split-half reliability estimation (see Fig.4, lower facet). Specifically, only Reaction Time (RT) and Efficiency demonstrate relatively better reliability, falling within the range of moderate. Specifically, under the "Self vs 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 "Self vs Stranger" contrast, the ICC2 for RT is 0.58 (95% CI = [.45, .73]), and for Efficiency, it is 0.35 (95% CI = [.21, .52]). Conversely, all other outcome variables exhibit reliability that is deemed unacceptable or poor.



**Fig. 4.** The weighted average split-half reliabilities (Monte-Carlo) and Intraclass Correlation Coefficient for different SPE measures.

*Note:* The vertical axis represents 12 different SPE measures, combining six indicators (RT, ACC, *d’*, *η*, *v*, *z*) and four baseline conditions (close other, stranger). The weighted average split-half reliability (upper figure) and ICC values and their corresponding 95% confidence intervals are illustrated using points and lines. The dashed line indicates that the confidence interval for that point estimate extends across 0, implying a non-significant value. Due to the fact that there is only one paper for 'Celebrity' and one for 'Nonperson,' their results is presented in the supplementary materials.

# **5 Discussion**

In this pre-registered study, our objective is to investigate the reliability of the different SPE measures in the SPMT. Our analysis reveals that, except parameters *z* from DDM, all the other measures exhibit robust SPE. However, when it comes to the reliability of SPE, it is acceptable or moderate at best for two indicators, Reaction Time and Efficiency, among all indicators that has been reported in the literature. Our results suggest that the current implementation of SPMT is not well-suited for assessing individual differences. Taken group-level and individual level result together, our study revealed a "reliability paradox" of SPE as measured by SPMT. These findings provide important methodological insights to future studies of SPE.

First, the reaction times and efficiency appear to be the best indicators among all different ways to measure SPE. Our results revealed that the reaction times and efficiency produced moderate to large effect sizes and comparably higher reliability. Therefore, for researchers interested in obtaining the group-level SPE as in the original experiment, reaction times and efficiency maybe the most sensitive indicators. Please note that the meta-analytical results reported here is a mini-meta-analysis for assessing the robustness of SPE with original experimental design of Sui et al (2012), not a meta-analysis based on a systematical literature search. For later, please see our on-going project (<https://osf.io/euqmf>). Additionally, we have conducted an exploratory analysis involving trial numbers and effect sizes, as detailed in the supplementary materials. The outcome of this analysis indicates insignificance.

Second, taken the group-level robustness and individual-level result together, our findings align with previously proposed “reliability paradox” in cognitive tasks (Hedge, Powell, & Sumner, 2018; Logie et al., 1996). We observed that the majority of the SPE measures demonstrated moderate to large effect sizes when analyzed at the group level. However, when delving into individual differences, only the SPE measures derived from reaction times (RT) and Efficiency displayed comparatively higher values than other SPE measures but still did not meet the criteria for satisfactory split-half reliability. Likewise, when examining reliability across multiple time points using ICC2, RT and Efficiency still ranked the higest but only showed moderate levels of test-retest reliability. Furthermore, our additional exploratory analysis on the relationship between split-half reliability and effect size (Hedges’ *g*) did not yield a significant result (refer to supplementary figure 6). These findings align well with the concept of the "reliability paradox," wherein robust experimental effects do not consistently correspond with robust correlations among individual differences (Logie et al., 1996). The reasons for reliability paradox have been extensively discussed in various papers (e.g., Hedge, Powell, Bompas, et al., 2018; Hedge, Powell, & Sumner, 2018; Kucina et al., 2023). SPMT, being a cognitive task, share many of of the underlying reasons, as elucidated in these previous works. Most importantly, traditional cognitive tasks, characterized by minimal variability between participants, hold particular value for researchers due to their near-universal ability to evoke the targeted effect (e.g., nearly all participants exibit SPE) (Liljequist et al., 2019). However, tasks that generate minimal between-participant variability exhibit limited capacity to discerningly rank individuals based on certain attributes, which conforms to the concept of reliability within the context of individual difference research (Hedge, Powell, & Sumner, 2018). Thus, it is important to note that the reliability should not be interpreted solely as a measure of the task's overall quality, but rather as an indication of the specific types of questions the SPMT can effectively address.

The relatively low reliability of all the SPE measures calls for attention when researchers are interested in measuring individual differences, such as in clinical settings (e.g., Karvelis et al., 2023), or searching an association with data from questionnaires (Hedge, Powell, & Sumner, 2018). The SPMT was originally designed to achieve robust group-level SPE rather than to measure individual differences. Consequently, if researchers are interested in assessing individual difference, there is a need for re-design the task so that it can capture performance nuances and reveal individual differences, as suggested by Kucina et al. (2023). For instance, increasing the number of trials may prove beneficial in enhancing reliability (please refer to the supplementary.1.4). Another possible solution is gamification, which have been shown to be able to improve data quality and assessment efficacy (Friehs et al., 2020). Data analytical methods might also improve the reliability. For instance, Enkavi et al. (2019) proposed that combing different cognitive tasks' data can capture the latent variables and increase the reliablity. Similarly, Haines et al. (2020) suggested generative models can be used as an alternative of ICC or test-retest reliaiblity.

Finally, another noteworthy result is the notably low split-half and test-retest reliability observed in the parameters (*v* and *z*) derived from the drift diffusion model. In our analysis, we applied common and easy-to-use methods to datasets that included at least 60 trials, and estimated parameter values for each condition of each participant and then calculated the reliability. The reliability of both the drift rate (*v*) and the starting point (*z*) fell well below acceptable levels. These findings raise concerns about the applicability of the standard cognitive model to effectively capture the underlying cognitive processes involved in the SPMT (Groulx et al., 2020). Similarly, a recent study by Schaaf et al. (2023) dicovered poor reliability for the parameter estimates of the standard reinforcement learning model in cognitive tasks. Consequently, it may not be advisable to apply the standard DDM directly to SPMT data without adhering to a comprehensive cognitive modeling workflow (e.g., Wilson and Collins (2019). Future research should therefore conduct a more thorough investigation of the DDM and explore the possibility of either adapting an existing model or developing a new one tailored to better capture the cognitive processes specific to the SPMT. Theoretically, this endeavor has the potential to enhance the reliability of these parameters.

## Implications of the current study

Our study yields practical implications for researchers aiming to use SPMT or improve it. First, for researchers who were grapple with determining which indicator is reliable, our analyses indicate that Reaction Time and Efficiency emerge as the most reliable measure of SPE, among all measures available in the SPMT. Additionally, our initial exploratory analysis on trial numbers suggests a potential link between low reliability and insufficient trial numbers. Therefore, we recommend incorporating more trials per condition, at least in line with the original SPMT, to enhance reliability. For researchers who interested in using SPMT as a measure of individual difference or with repetaed adminstration, the current version of SPMT demonstrates limitations in capturing trait-like characteristics. Consequently, significant modifications are imperative.

## Limitations

Several limitations warrant acknowledgment. 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 "(W)EIRD" populations (Rad et al., 2018; Yue et al., 2023). 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 (ICC2), only one dataset had longitudinal data available, which could potentially limit the representativeness of the results. This issue is mitigated by the fact that additional analysis of two datasets (see suplemetary) that with different design showed similar results as we reported in the main text. 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.

# **6 Conclusion**

This study provides an empictal assessment of the reliability of the self-perceptual matching task (SPMT). We found a robust self-prioritization effect for reaction times and efficiency. Mean while, the reliability of the most robust SPE measure falls short of being satisfactory. The results of the current study may serve as a bench marker for the improvement of future studies.**Acknowledgements**

The authors declare that this research received no external funding.

# **Author contributions**

HCP contributed to the conception and supervision of the study. HCP contributed to data collection. ZL, ZYR and HMZ 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. HCP, JS, ZL and HMZ 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**

## Parameter Recovery Result for Package Comparison

We chose not to utilize the HDDM package (Wiecki et al., 2013) since the computation process is significantly time-consuming, necessitating high computational resources and leading to prolonged overall analysis time. Instead, we performed a package comparison by generating 100 datasets using the HDDM package in Python, in order to identify the most appropriate package for our analysis. These datasets were specifically configured with parameters *a* = 2, *t* = 0.3, *v* = 1, and *z* = 0.7. Subsequently, we utilized three different DDM packages in R, namely RWiener (Viechtbauer, 2010), hausekeep (Lin, 2019), and FastDMinR (Voss & Voss, 2007), to compute parameter estimates for these generated datasets. The evaluation process involves comparing the computed values obtained from the R packages with the set parameters. 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.

Figure 1 presents the results of the package comparison. The estimated drift rate (*v*) obtained from RWiener is 1.01, with a 95% confidence interval of [.98, 1.03], which closely aligns with our pre-defined values. Similarly, the estimated starting point (*z*) is 0.77, with a 95% confidence interval of [.76, .78], also very close to our pre-defined value. On the contrary, the parameters calculated using other packages either showed high inaccuracies, excessively wide confidence intervals, or required extended computation times. As a result, we have opted to utilize RWiener for our calculations. It strikes a favorable balance between accuracy, confidence interval width, and computational efficiency, making it the most suitable choice for our analysis.



**Supplementary Fig. 1 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.

* 1. **Split-half Reliabilities Using Four Splitting approaches.**

In this section, we are presenting the split-half reliabilities results for the SPE measures using four split-half methods: Monte Carlo, first-second, odd-even, and permutated. We also include includes the drift rate (*v*) and starting point (*z*) estimated from the "hausekeep" package in the analysis. However, it is important to highlight that the estimation of parameter "*a*" in "hausekeep" significantly deviates from the HDDM approach, primarily because of the assumption that *z* = *a* / 2 (refer to suppl. Fig. 1). As a result, we have chosen not to include the results obtained from this package in the main text. Nevertheless, we present them here for reference and transparency. Please refer to Fig. 2(a-d) for the visual representation of the results.

A graph with black lines and white text

Description automatically generated

**Supplementary Fig. 2a Results of Split-half Reliabilities using Monte Carlo Split-half Methods.**

A graph with black lines and white text

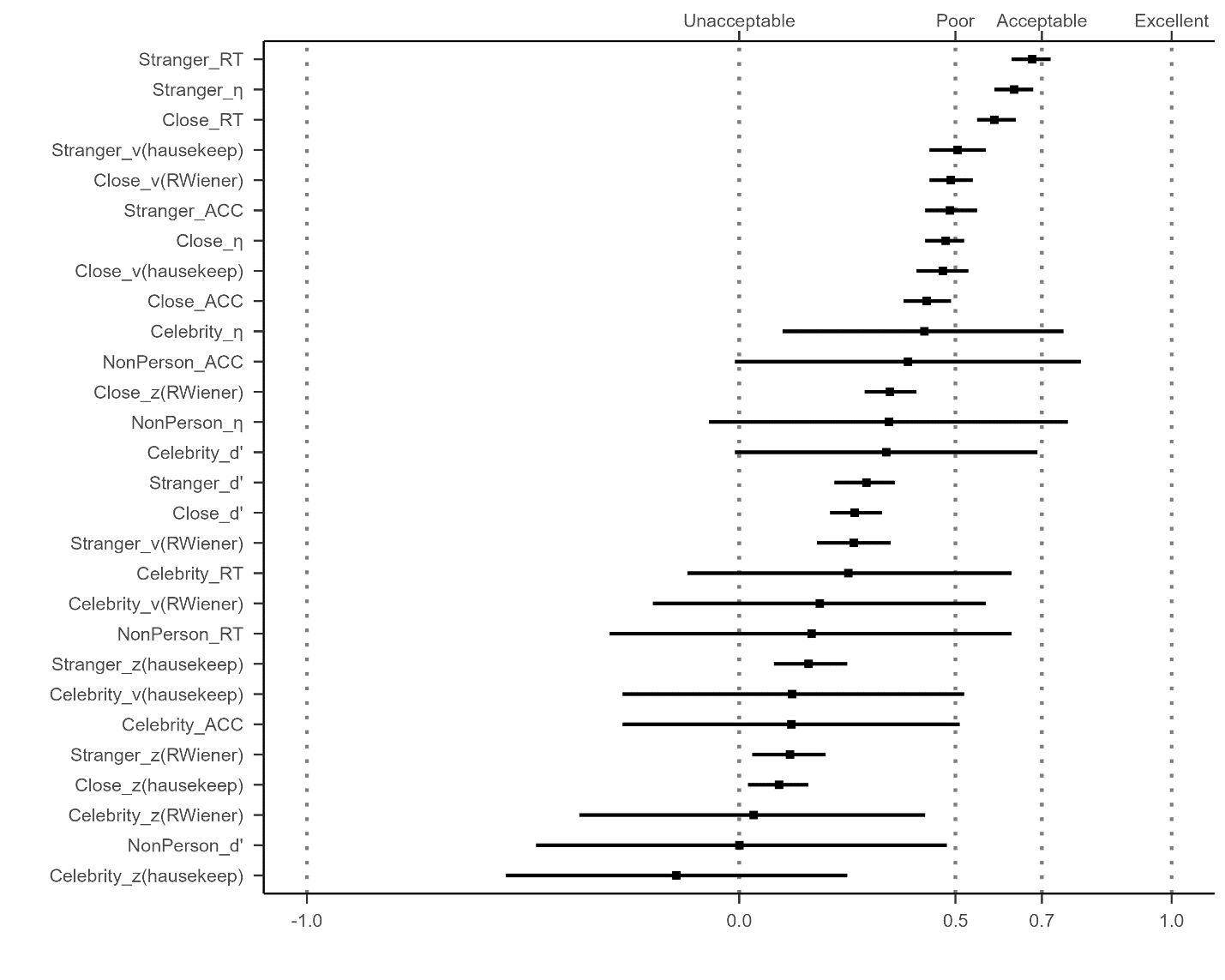
Description automatically generated

**Supplementary Fig. 2b Results of Split-half Reliabilities using First-Second Split-half Methods.**

A graph with black lines and white text

Description automatically generated

**Supplementary Fig. 2c** **Results of Split-half Reliabilities using Odd-Even Split-half Methods.**



**Supplementary Fig. 2d** **Results of Split-half Reliabilities using Permuted Split-half Methods.**

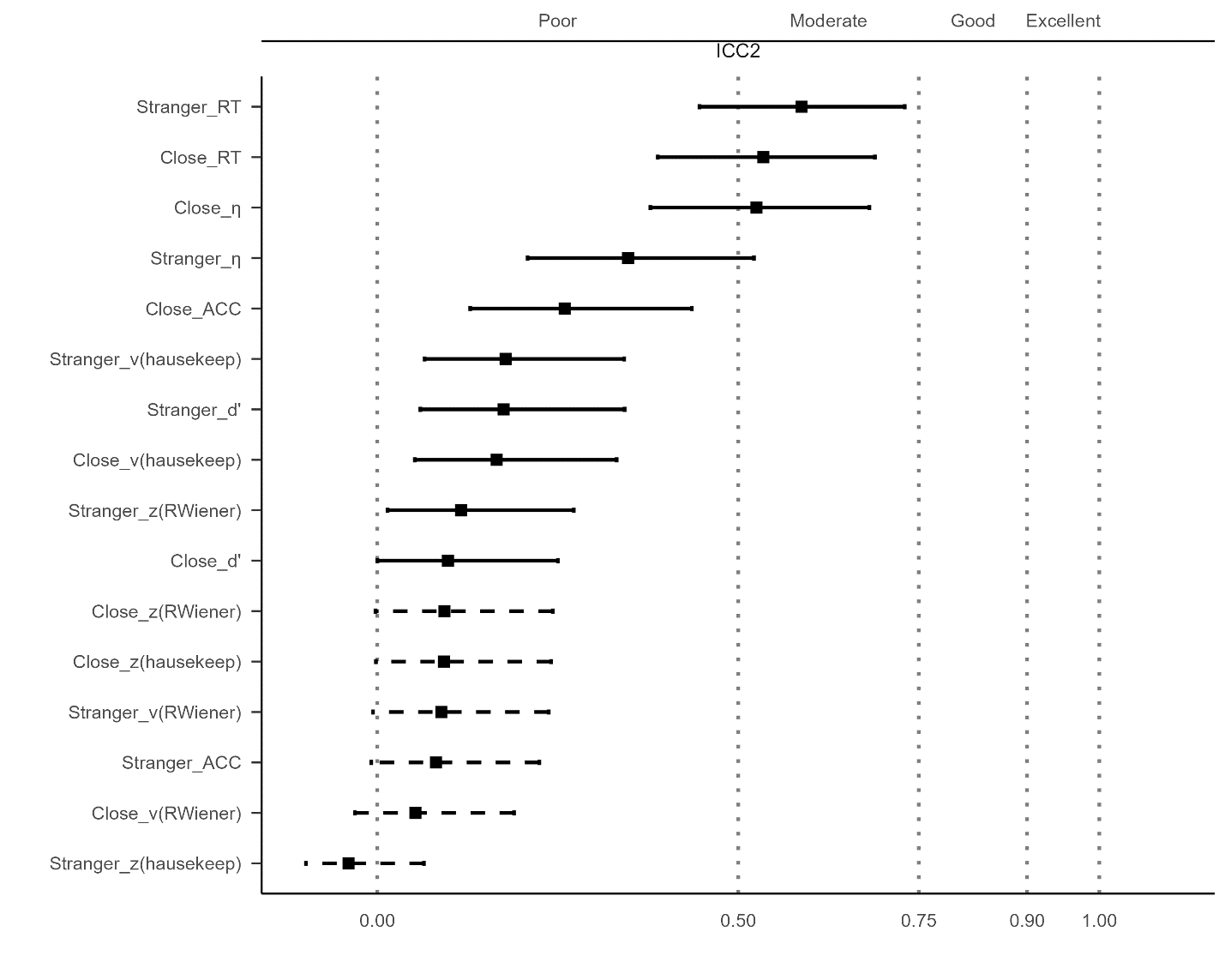
*Note:* The vertical axis of the graph listed 32 different SPE measures, combining six indicators (RT, ACC, *d’, η, v, z*) and four baseline conditions (close other, stranger, celebrity, and non-person). The *v* and *z* implemented using the "hausekeep" package were also included. The weighted average split-half reliability and 95% confidence intervals are shown by points and lines. The figure is divided into separate facets arranged from left to right, each representing weighted average split-half reliability calculated using three distinct methods: first-second, odd-even, and permuted.

It is evident that the results pattern from the permuted split-half methods and first-second split-half methods closely resemble the Monte Carlo method's outcomes. The top four split-half reliabilities, ranked highest, are as follows: Reaction Time (RT) with the "Stranger" contrast, Efficiency (*η*) with the "Stranger" contrast, RT with the "Close other" contrast, *η* with the "Self vs Close" contrast. However, the results obtained from the odd-even split-half method were notably different from the other three methods. We hypothesize that this discrepancy may be attributed to the odd-even method's sensitivity to temporal dependencies, which could have been influenced by the inherent sequential nature of responses in the SPMT. Further investigation into the presence and impact of serial dependency in the data would be valuable to better understand the observed variations in the split-half reliabilities among the different methods.

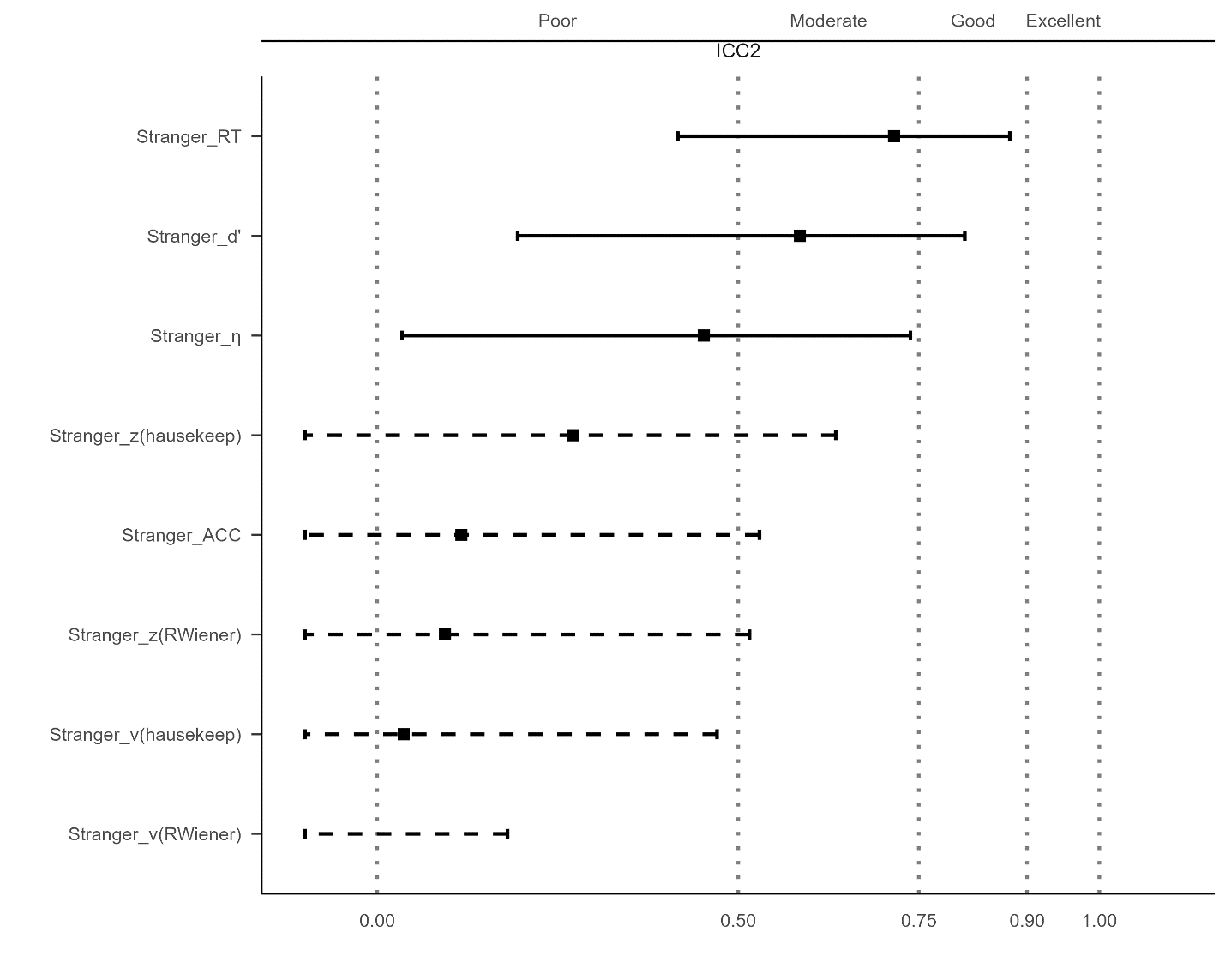
## ICCs for SPE Measures Using Other Dataset

In Suppl. Fig. 3a, we present the results of the Intraclass Correlation Coefficients (ICC2) for the SPE measures, where drift rate (*v*) and starting point (*z*) estimated from the "hausekeep" package were also included. In Fig. 3b, we extended our exploration of ICC2 to include the SPE measures from one additional dataset. However, it is important to note that the SPMT used in this dataset deviated quite strongly from the original SPMT paradigm. Due to these significant differences, ICC2 obtained from this dataset may reflect variations introduced by the modified SPMT rather than directly comparable results to the original paradigm.

In addition, in Fig. 3c, we included the BDI-II scores of each participant as covariates in the calculation of ICC. We found that the ICC2 remained relatively consistent before and after taking BDI-II scores into account as a covariate. This may be due to small participant numbers, out of 33 participants, only six had high BDI scores (above 20). Therefore, even though the ICC did not significantly change after considering BDI-II scores, this result does not necessarily imply that depression has no impact on SPMT outcomes.

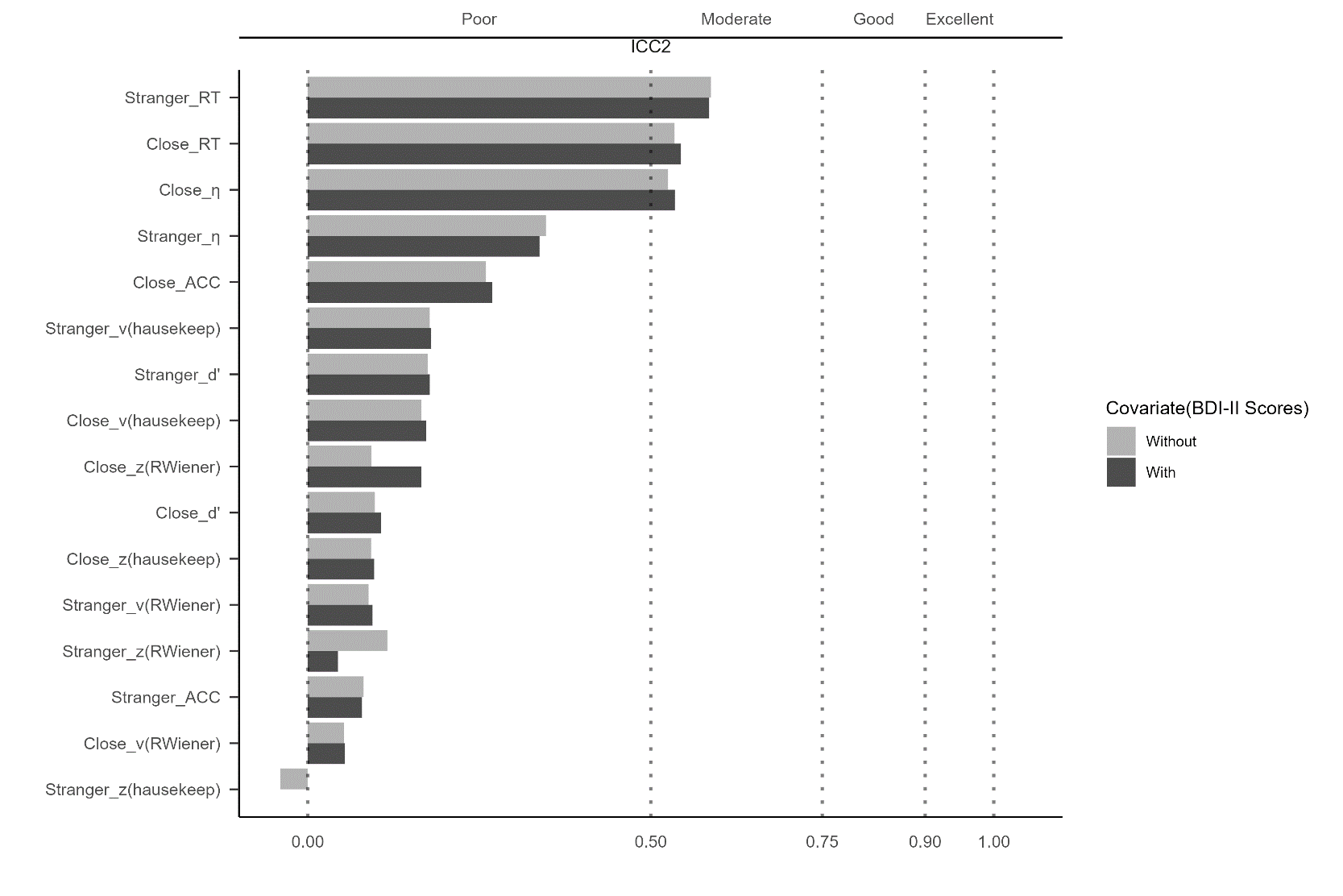


**Supplementary Fig. 3a ICC2 for SPE Measures Using Hu et al. (2023).**



**Supplementary Fig. 3b ICC2 for SPE Measures Using an Aditional dataset.**

*Note:*The vertical axis of the graph illustrates eight distinct indicators, which includes two additional indices from the DDM, implemented using the "hausekeep" package. The line and dots on the graph represent the value of ICC2, along with their corresponding 95% confidence intervals. The dashed line indicates that the confidence interval for that point estimate extends beyond the range of our coordinate axes (0, 1).

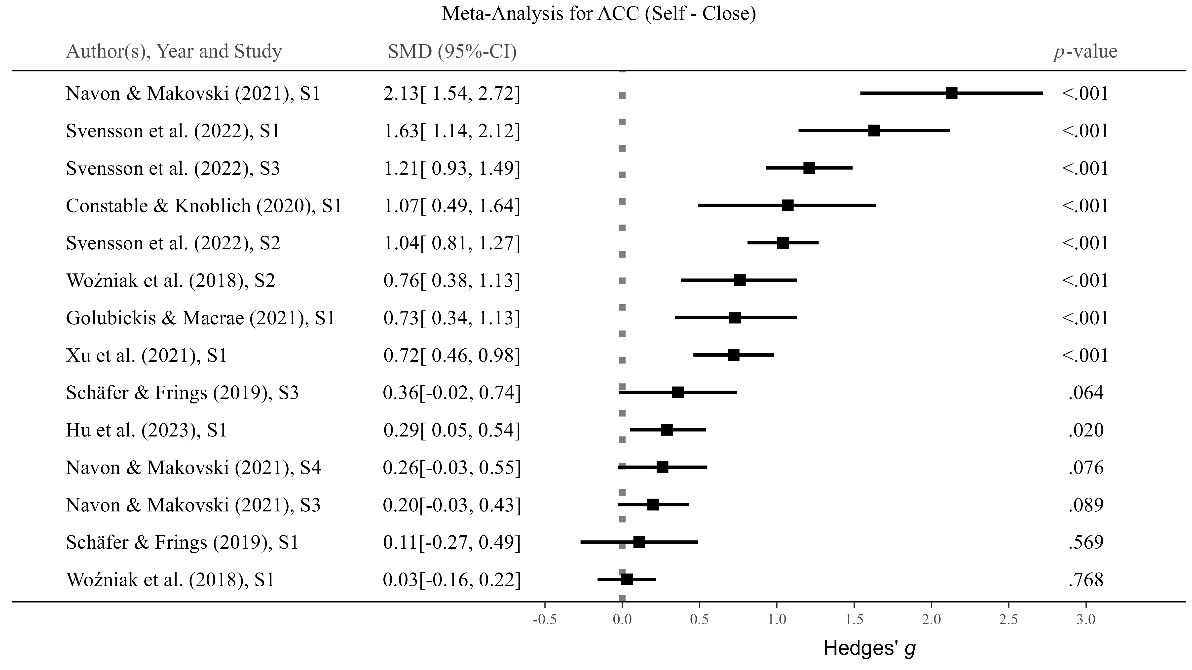


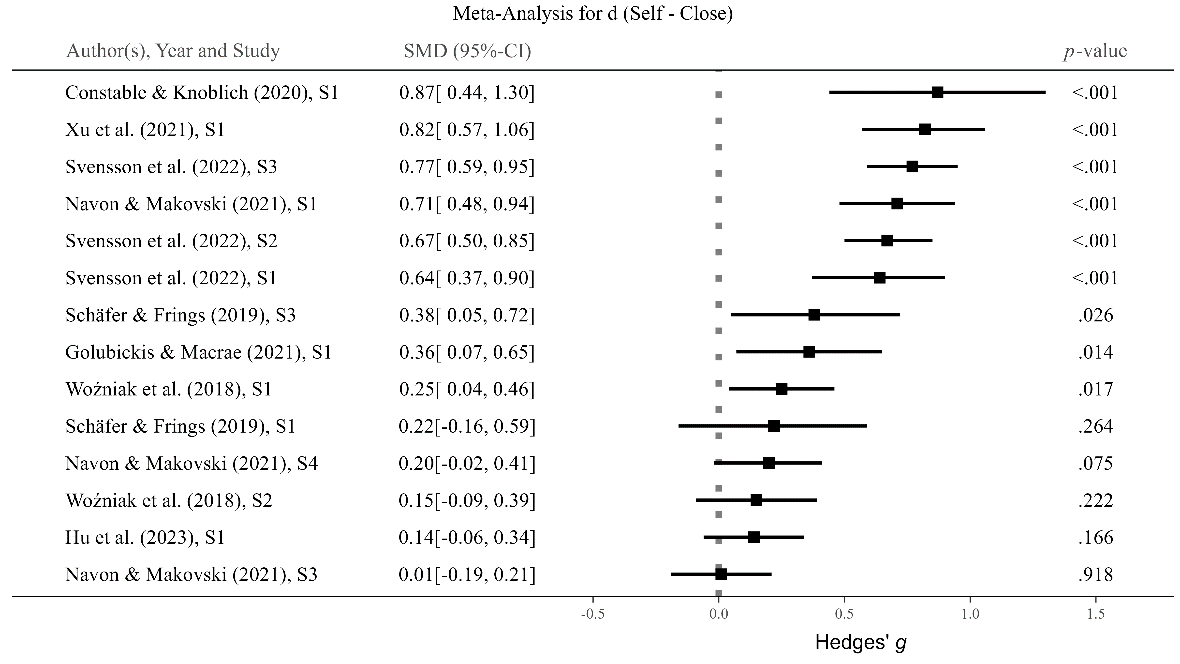
**Supplementary Fig. 3c ICC2 for SPE Measures Using Hu et al. (2023) with and without covariate (BDI-II Scores).**

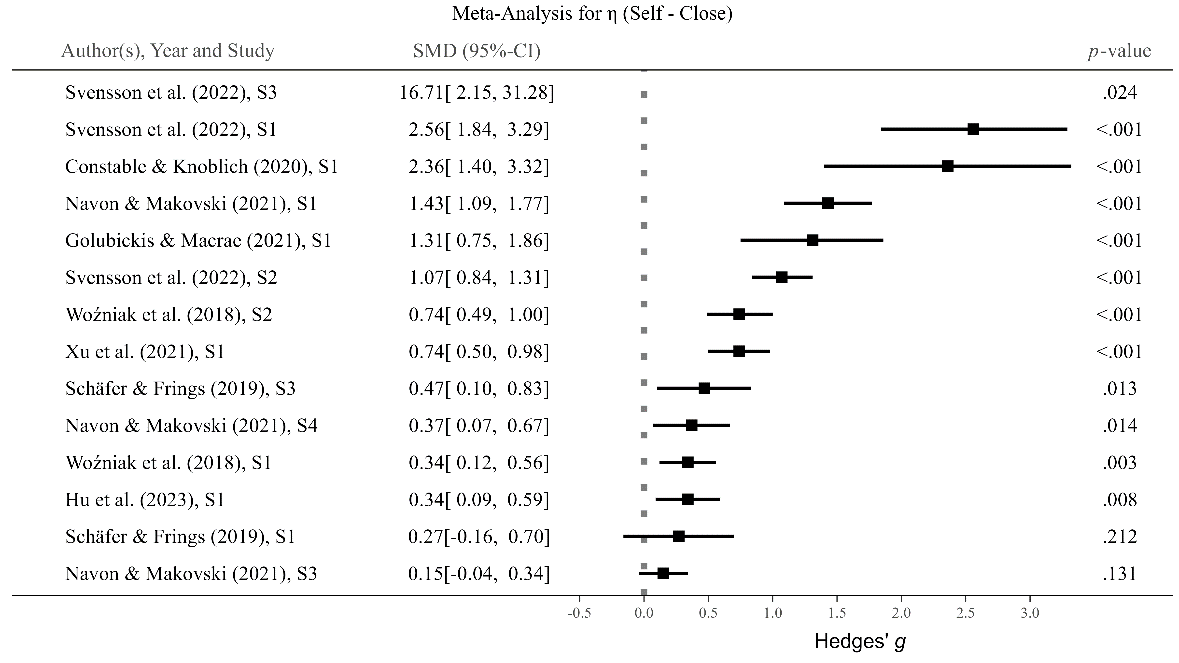
*Note:*The vertical axis of the graph illustrates eight distinct indicators, which includes two additional indices from the DDM, implemented using the "hausekeep" package. The bar on the graph represent the value of ICC2.

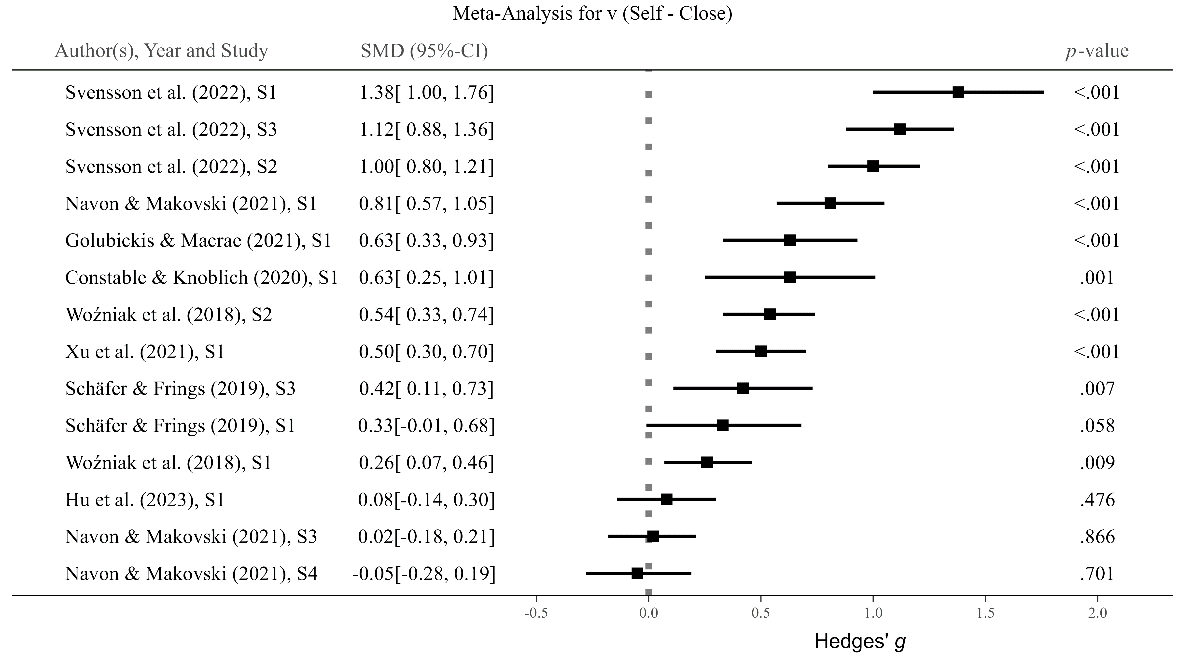
## Meta Analysis Results

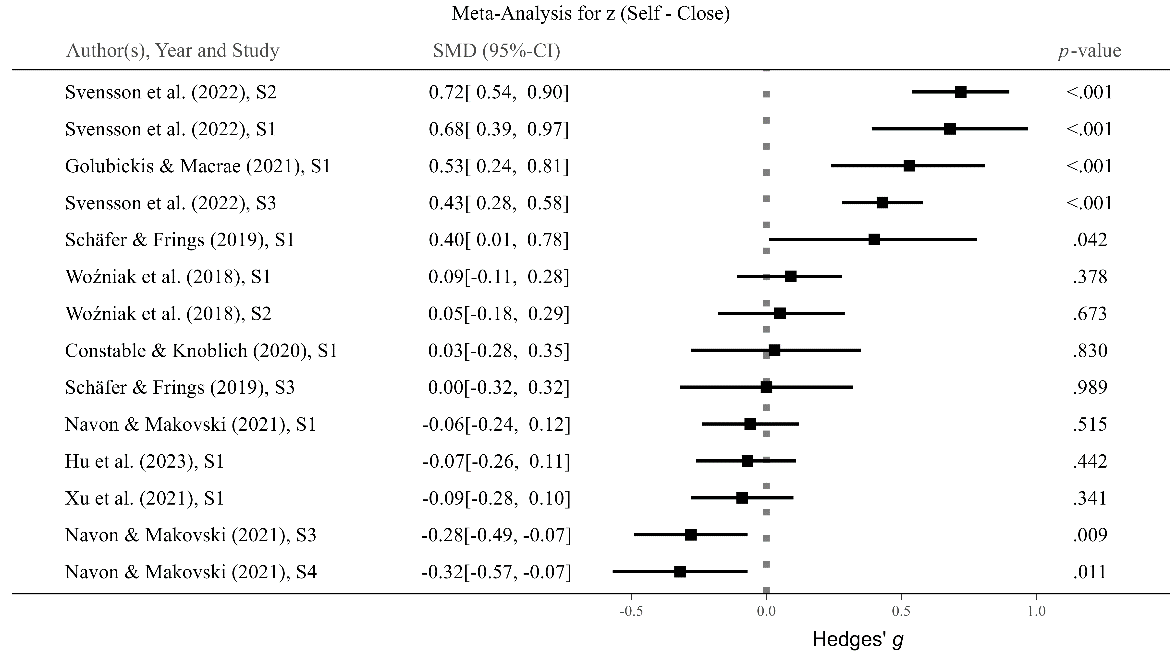
We conducted a meta-analysis of all six SPE measures. These Forest Plots are shown in Fig. 4.

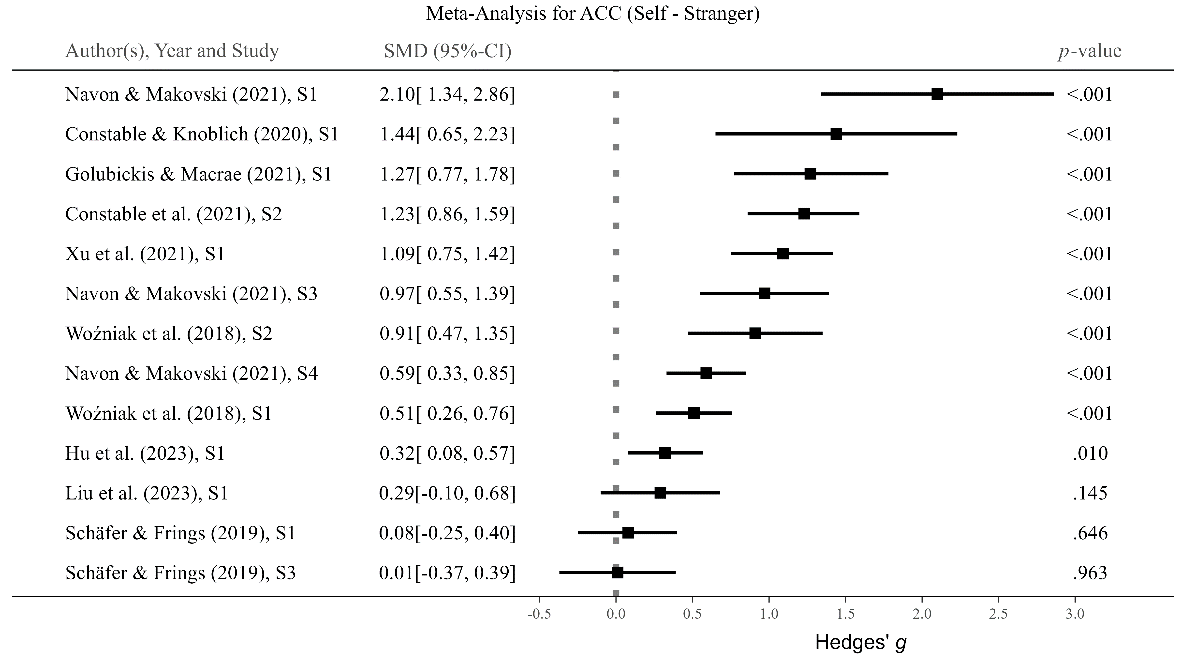


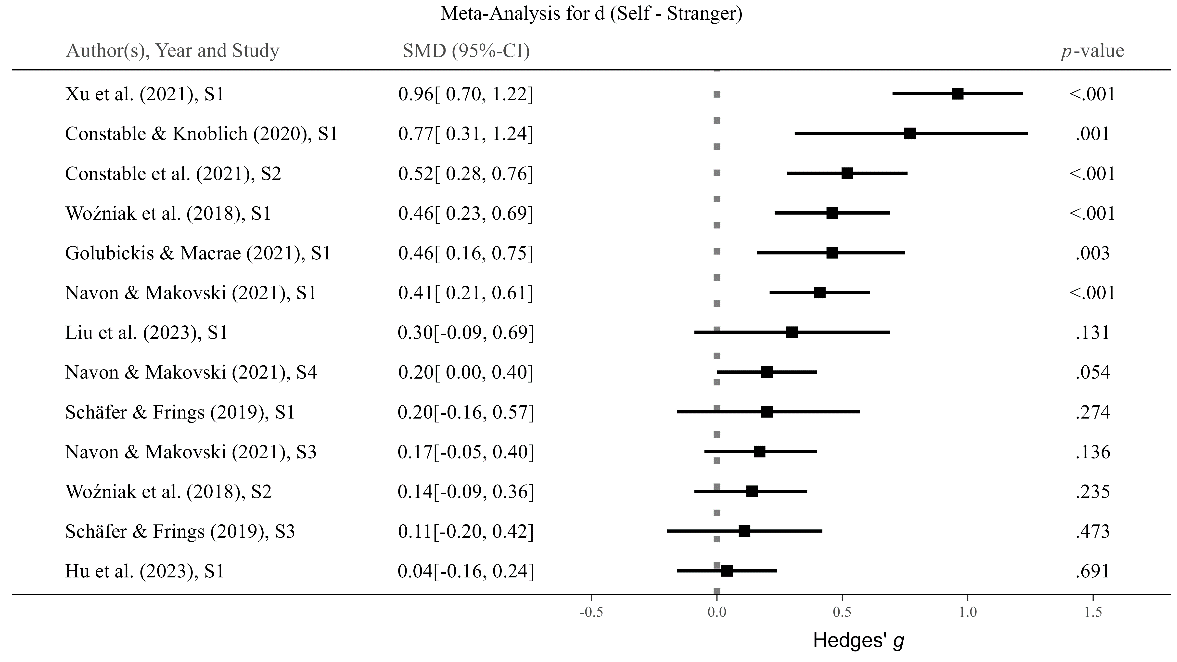


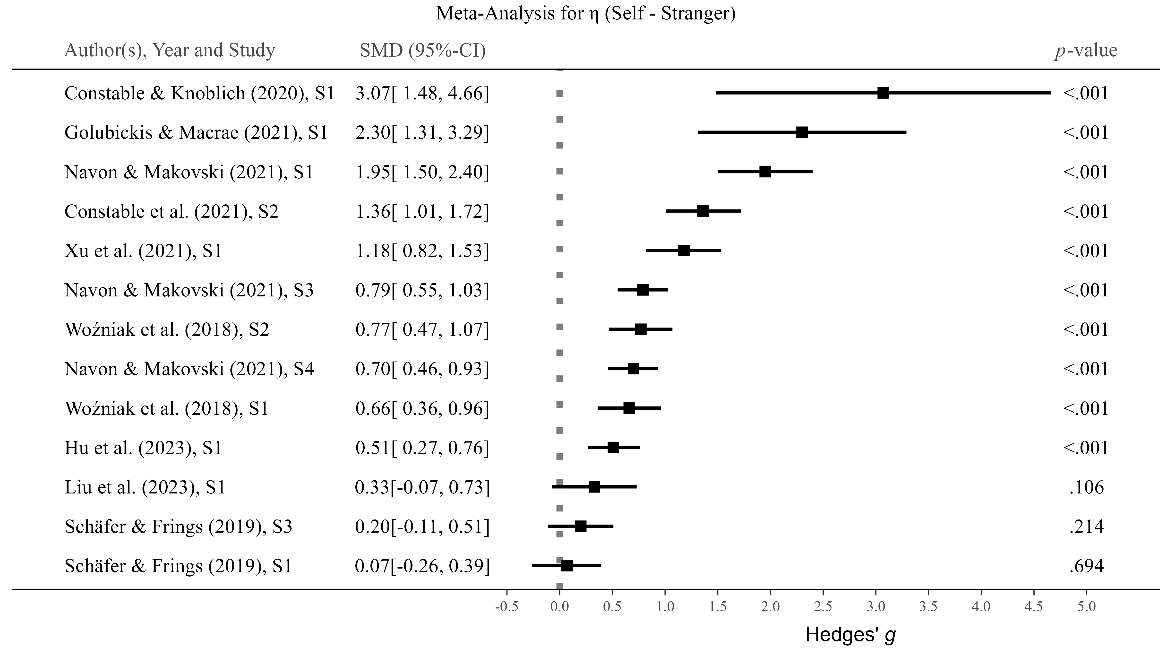


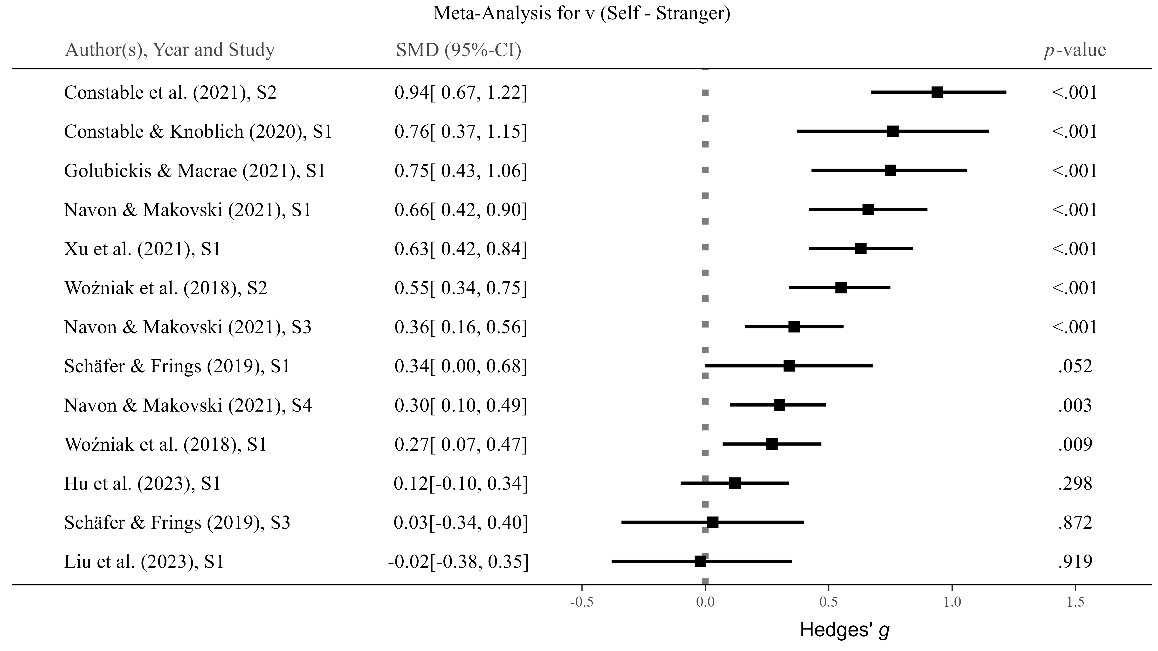


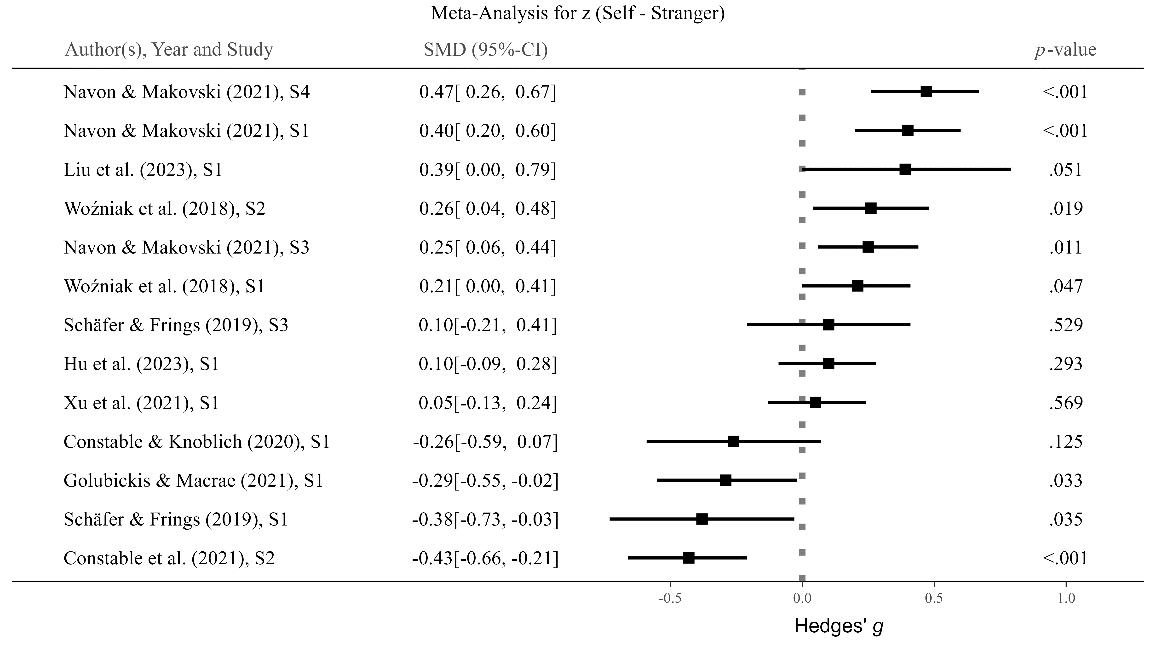












**Supplementary Fig. 4 Forest Plot of indicators of SPE**

## Exploratory analysis

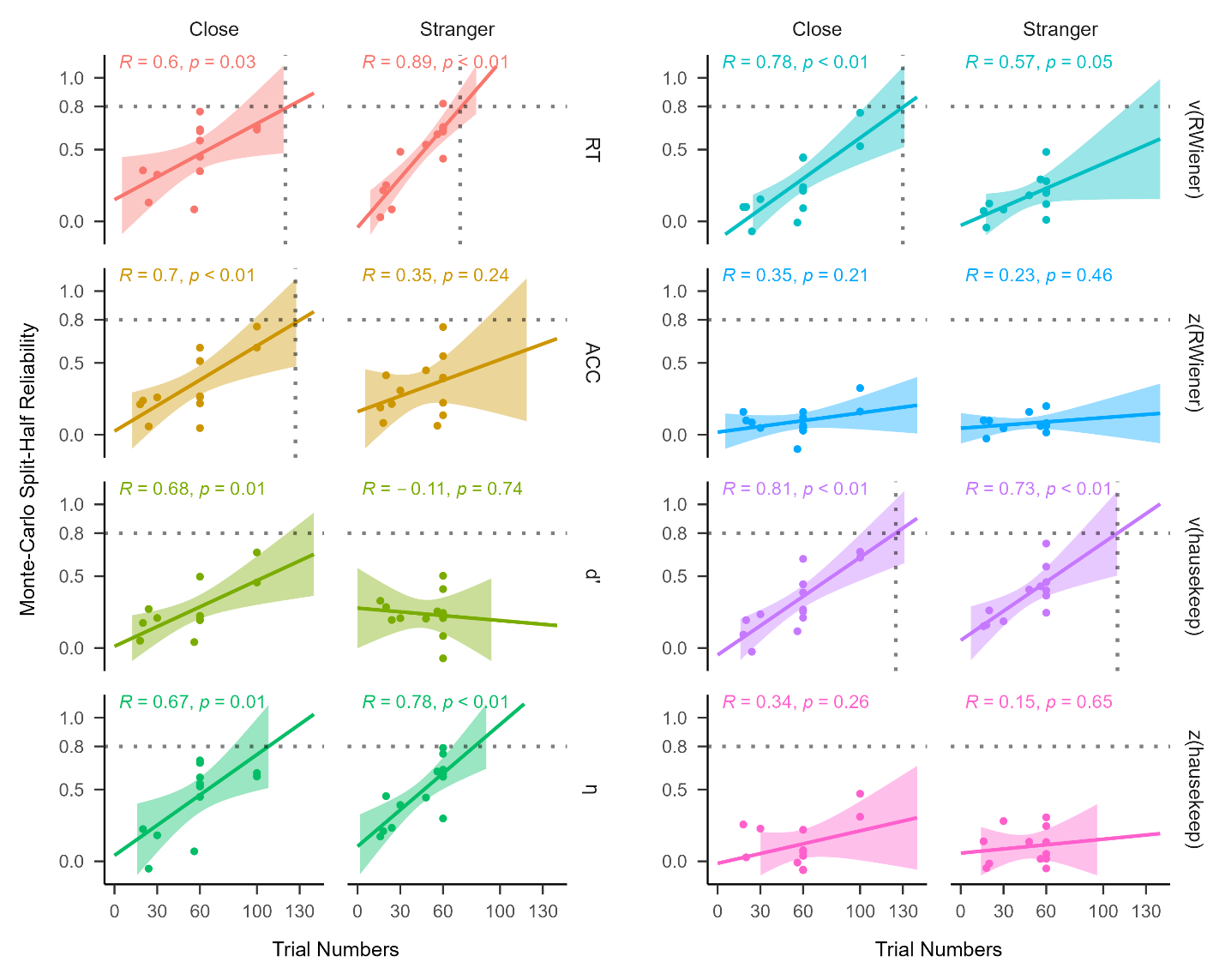
In this section, we present the results of the correlation analysis between Monte Carlo split-half reliability, the number of trials and the effect size (Hedges’ *g*). Notably, we found significant correlations between Monte Carlo split-half reliability and trial numbers for some indicators, such as Reaction Time and Efficiency. However, for indicators like *d'* and *v*, the correlation with trial numbers was relatively weak. However, the effect size (Hedges’ *g*) exhibits generally low correlations with Monte Carlo split-half reliability and trial numbers.

From the Supplementary Fig. 5, we can observe that the SPMT paradigm requires approximately 80 trials to achieve a Monte Carlo split-half reliability of 0.8 for the SPE measure of RT under the 'Stranger' condition and around 120 trials under the 'Close' condition. Furthermore, achieving a Monte Carlo split-half reliability of 0.8 for the *v* parameter may require more than 120 trials. On the other hand, attaining high Monte Carlo SHR values for the remaining three indicators, particularly for the *z* parameter, remains challenging even with 150 or more trials.

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 at least 80 trials under each experimental condition.

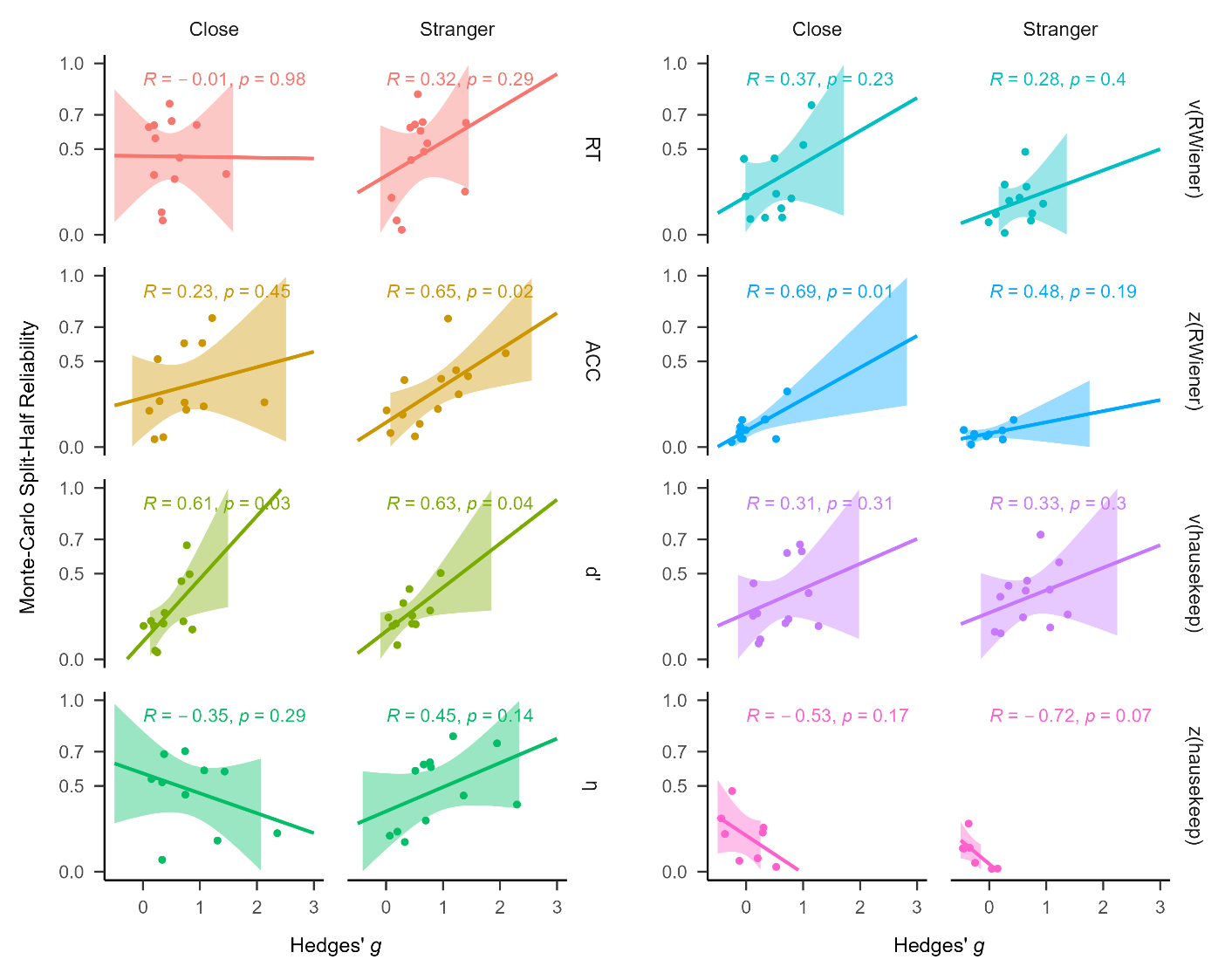
As Kucina et al. (2023) pointed out, the larger the effect size between the experimental and control groups, the higher the split-half reliability tends to be. Therefore, we explore the correlation between split-half reliability and effect size (Hedges’ g), as shown in Supplementary Fig. 6. Our exploratory analysis did not find a high correlation among them. This result pattern is some how consistent with the reliability paradox proposed by Logie et al. (1996), which suggests that robust experimental effects are not always associated with robust individual difference correlations.

At last, we also calculated the correlation coefficient between trial numbers and effect size (Hedges’ g), as shown in Fig. 7. Similarly, no significant correlation was found.



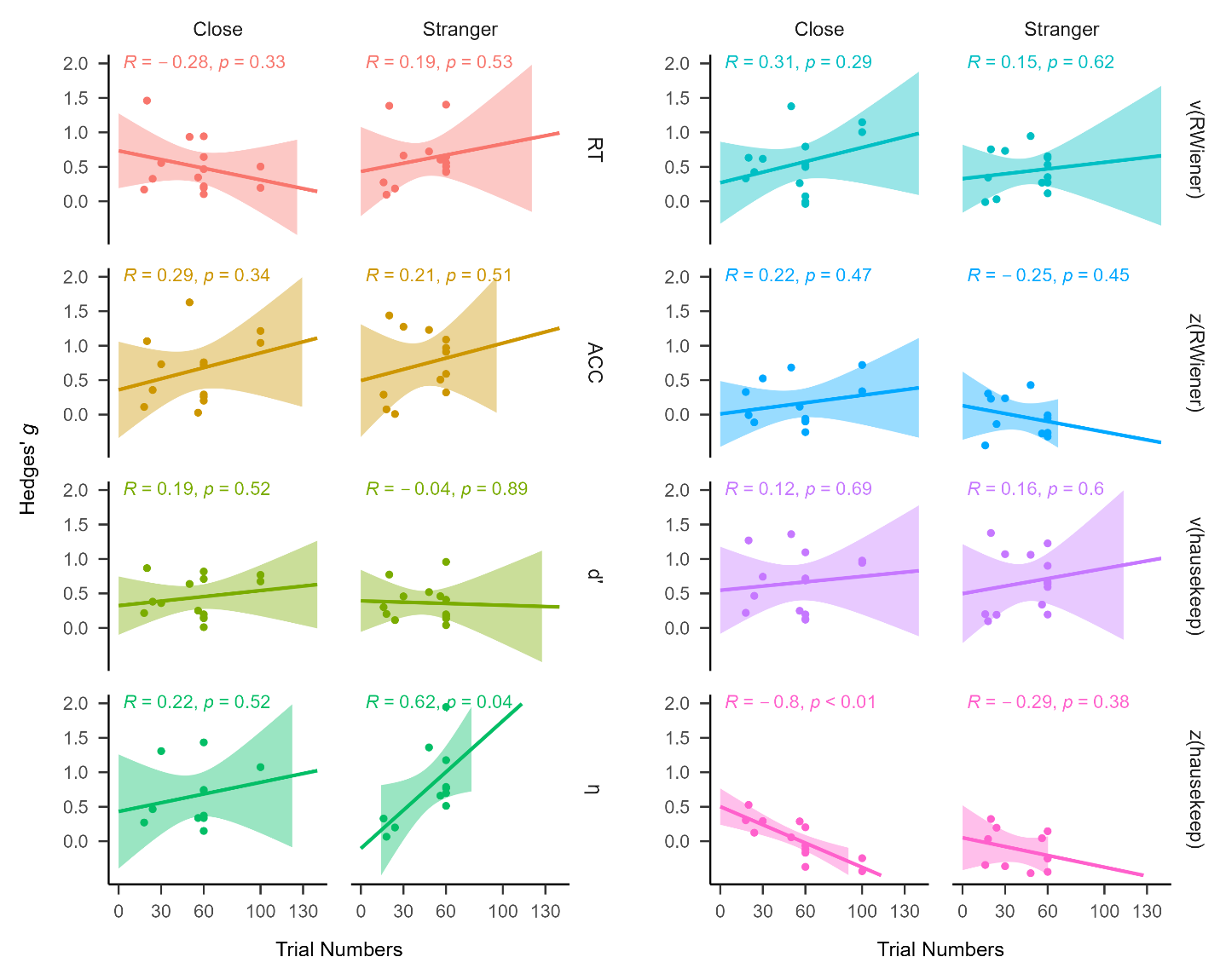
**Supplementary Fig. 5 Regression analysis between Monte Carlo split-half reliability and trial numbers using different SPE measures.**

*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.



**Supplementary Fig. 6 Regression analysis between Monte Carlo split-half reliability and effect size (Hedges’ *g*) using different SPE measures.**

*Note:* The vertical axis represents Monte-Carlo split-half reliability, and the horizontal axis represents the effect size (Hedges’ g). Each facet represents one SPE measures.



**Supplementary Fig. 7 Regression analysis between Trial Numbers and effect size (Hedges’ *g*) using different SPE measures.**

*Note:* The vertical axis represents the effect size (Hedges’ g), and the horizontal axis represents trial numbers. Each facet represents one SPE measures.

## Experimental design (Hu et al., 2023).

### Participants

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

36 participants were recruited from Tsinghua University and nearby community, all of whom were right-handed and had normal or corrected-to-normal vision. Participants were pre-tested for their depressive level by Beck Depression Inventory (BDI). Only participants with score lower than 10 (n=30) and those higher than 20 (n=6, ID: 6005, 6006, 6010, 6022, 6025, and 6036) were included. Data from three participants (ID: 6008, 6015, and 6031) should be excluded due to invalid trials or program malfunctions. The exclusion left 33 valid participants (Mean Age = 21.06, SD Age = 3.24), with 21 females and 12 males.

### Experimental design

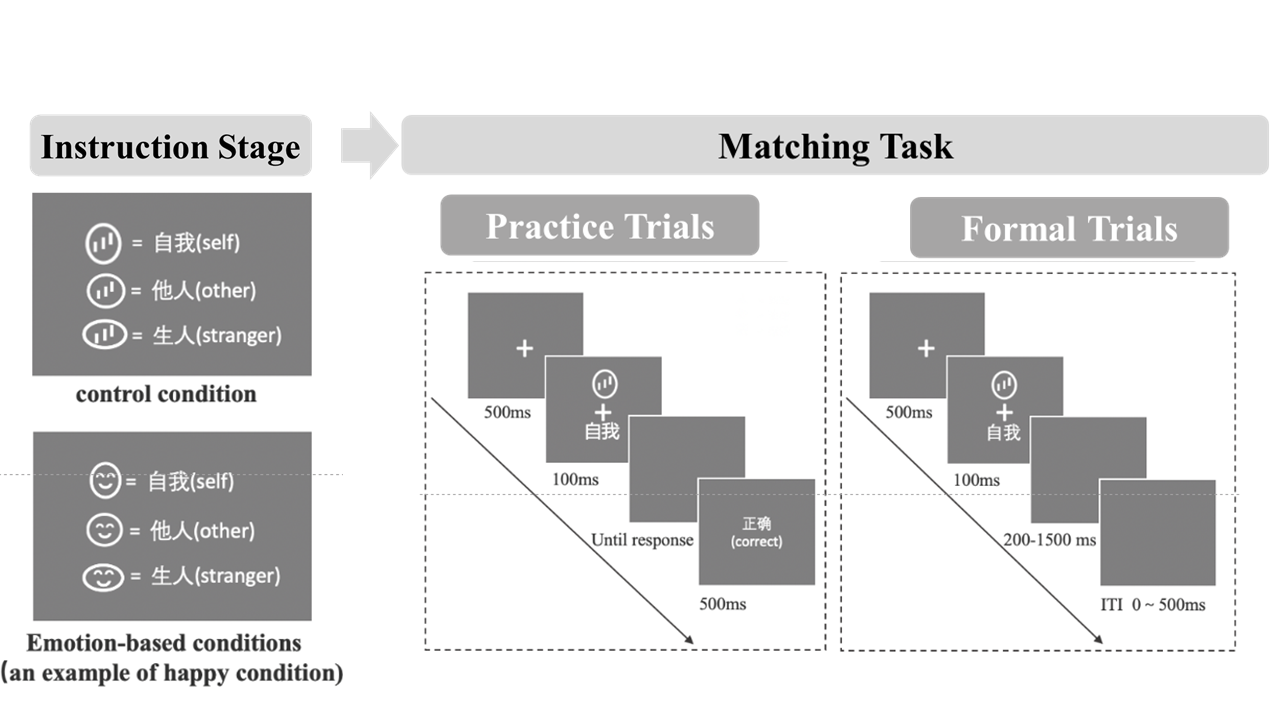
Experiment 2 is a 2 (Matching: Matching vs. Nonmatching) × 3 (Idnetity: Self, Friend, Stranger) × 4 (Emotion: Control, Neutral, Happy, Sad) × 6 (Sessions: 1-6) experiment.

### Procedure

The experiment was finished individually in a dimly lighted room. Stimuli were presented and responses were collected using E-Prime 2.0 on PC. The monitor was at 1024 × 768 resolution with 100 Hz refresh rate.

The experiment had two phases. Following Sui et al. (2012), the first phase comprised a instruction stage in which participants were required to associate geometric shapes with labels. The shapes were not presented at this stage. The instruction stage lasted for approximately 60 seconds and shape-target associations were counterbalanced across the sample. Next, participants performed a matching task. At the start of each trial, a fixation cross was first displayed in the center of the screen for 500 ms. Then, a shape–label pairing as well as the fixation cross was presented for 100ms, respectively. The next frame showed a blank screen for 1500 ms, or until a response was made. Participants were asked to determine whether the shape was appropriately matched to the label by pressing one of the two response buttons as quickly and precisely as possible within this timeframe.

The participants need to separately learn 4 sets of association between shapes and labels. The associations contain 1 control condition and 3 sets of emotion-based condition. In the control condition, participants learned the association between 3 geometric shapes (circle, horizontal ellipse and vertical ellipse) and three labels (Self, Friend, Stranger). In each of the emotion-based condition, participants would see facial expressions (happy, sad, neutral) appear on the circle, horizontal ellipse and vertical ellipse (see Supplementary Fig. 8). In each condition, before commencing the formal experimental trials, participants underwent a training session comprising 24 practice trials. After the practice trials, each participant completed 6 blocks of 60 trials in the task. There were six types of shape-label associations: 2 Matching (Matching / Nonmatching) x 3 Shape (Self, Friend, Stranger) associations, with 60 trials for each association. Participants took a short break (up to 60 seconds) after each block. Each participant was required to repeat the experiment six times, with a one-week gap between each wave of experiments.



**Supplementary Fig. 8** Procedure of the SPMT in the Experiment 2 (Hu et al., 2023)..

*Note:* Examples of stimuli and time course of the experimental procedure in Experiment. The labels and feedback appeared in Chinese in the experiment. In the associative learning task, the matched associations of shapes and labels was counterbalanced between participants. Timely feedback was not provided in formal trials.

# **References**

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

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

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

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

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

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

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic, 86*(2), 127-137. <https://psycnet.apa.org/record/1982-00095-001>

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

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>

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

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.

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.

Groulx, J. T., Harding, B., & Cousineau, D. (2020). The EZ Diffusion Model: An overview with derivation, software, and an application to the Same-Different task. *The Quantitative Methods for Psychology, 16*(2), 154-174.

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2020). Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox.

Hedge, C., Powell, G., Bompas, A., Vivian-Griffiths, S., & Sumner, P. (2018). Low and variable correlation between reaction time costs and accuracy costs explained by accumulation models: Meta-analysis and simulations. *Psychological Bulletin, 144*(11), 1200-1227. <https://doi.org/10.1037/bul0000164>

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.

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

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

Hu, C.-P., 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>

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

Kahveci, S., Bathke, A., & Blechert, J. (2022). Reliability of reaction time tasks: how should it be computed? <https://doi.org/10.31234/osf.io/ta59r>

Karvelis, P., Paulus, M. P., & Diaconescu, A. O. (2023). Individual differences in computational psychiatry: a review of current challenges. *Neuroscience & Biobehavioral Reviews*, 105137.

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>

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

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

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine, 15*(2), 155-163. <https://doi.org/10.1016/j.jcm.2016.02.012>

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

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. (2019). How to use hausekeep. <https://doi.org/10.5281/zenodo.2555874>

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.

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

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

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

Orellana-Corrales, G., Matschke, C., & Wesslein, A.-K. 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>

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Journal of clinical epidemiology, 134*, 178-189. <https://doi.org/10.1016/j.jclinepi.2021.03.001>

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.

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

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review, 29*(1), 44-54. <https://doi.org/10.3758/s13423-021-01948-3>

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

R Core Team. (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.

Revelle, W. R. (2017). psych: Procedures for personality and psychological research. <https://CRAN.R-project.org/package=psych>

Schaaf, J. V., Weidinger, L., Molleman, L., & van den Bos, W. (2023, 2023/09/08). Test–retest reliability of reinforcement learning parameters. *Behavior research methods*. <https://doi.org/10.3758/s13428-023-02203-4>

Schäfer, S., & Frings, C. (2019). Understanding self-prioritisation: The prioritisation of self-relevant stimuli and its relation to the individual self-esteem. *Journal of cognitive psychology, 31*(8), 813-824. <https://doi.org/10.1080/20445911.2019.1686393>

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

Sui, J., He, X., & Humphreys, G. W. (2012). Perceptual effects of social salience: Evidence from self-prioritization effects on perceptual matching. *Journal of experimental psychology. Human perception and performance, 38*(5), 1105-1117. <https://doi.org/10.1037/a0029792>

Svensson, S. L., Golubickis, M., Maclean, H., Falbén, J. K., Persson, L. M., Tsamadi, D., Caughey, S., Sahraie, A., & Macrae, C. N. (2022). More or less of me and you: self-relevance augments the effects of item probability on stimulus prioritization. *Psychological Research, 86*(4), 1145-1164. <https://doi.org/10.1007/s00426-021-01562-x>

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

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of statistical software, 36*, 1-48.

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39*(4), 767-775.

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. <https://doi.org/10.3389/fninf.2013.00014>

Wilson, R. C., & Collins, A. G. E. (2019, 2019/11/26). Ten simple rules for the computational modeling of behavioral data. *eLife, 8*, e49547. <https://doi.org/10.7554/eLife.49547>

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

Xu, T., Kiar, G., Cho, J. W., Bridgeford, E. W., Nikolaidis, A., Vogelstein, J. T., & Milham, M. P. (2023, 2023/07/01). ReX: an integrative tool for quantifying and optimizing measurement reliability for the study of individual differences. *Nature Methods, 20*(7), 1025-1028. <https://doi.org/10.1038/s41592-023-01901-3>

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>

Yue, L., Zuo, X., & Chuan-Peng, H. (2023). The WEIRD problem in a “non- WEIRD” context: A meta-research on the representativeness of human subjects in Chinese psychological research. <https://doi.org/https://doi.org/10.17605/OSF.IO/MTR8D>

Zhang, Y., Wang, F., & Sui, J. (2023, 2023/08/01/). Decoding individual differences in self-prioritization from the resting-state functional connectome. *NeuroImage, 276*, 120205. <https://doi.org/https://doi.org/10.1016/j.neuroimage.2023.120205>

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