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

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

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

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

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

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

1 Supplementary Methods

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

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

1.1.1 Ethics Information

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

1.1.2 Participants.

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

1.1.3 Experimental Design

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

1.1.4 Procedure

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

task. At the start of each trial, a fixation cross was first displayed in the center of the screen for 500 ms. Then, a shape–label pairing as well as the fixation cross were presented for 100ms, respectively. The next frame showed a blank screen for 800-1200 ms. Participants were asked to determine whether the shape was appropriately matched to the label by pressing one of the two response buttons as quickly and precisely as possible within this timeframe.
The participants needed to separately learn 4 sets of associations between shapes and labels. The associations contained 1 control condition and 3 sets of emotion-based conditions. In the control condition, participants learned the association between 3 geometric shapes (circle, horizontal ellipse and vertical ellipse) and three labels (Self, Friend, Stranger). In each of the emotion-based conditions, participants would see facial expressions (happy, sad, neutral) appear on the circle, horizontal ellipse and vertical ellipse (see Fig. S1). In each condition, before commencing the formal experimental trials, participants underwent a training session comprising 24 practice trials. After the practice trials, each participant completed 6 blocks of 60 trials in the task. There were six types of shape-label associations: Matching (Matching / Non-matching) x Shape (Self, Friend, Stranger) associations, with 60 trials for each association. Participants took a short break (up to 60 seconds) after each block. Each participant was required to repeat the experiment six times, with a one-week gap between each wave of experiments.

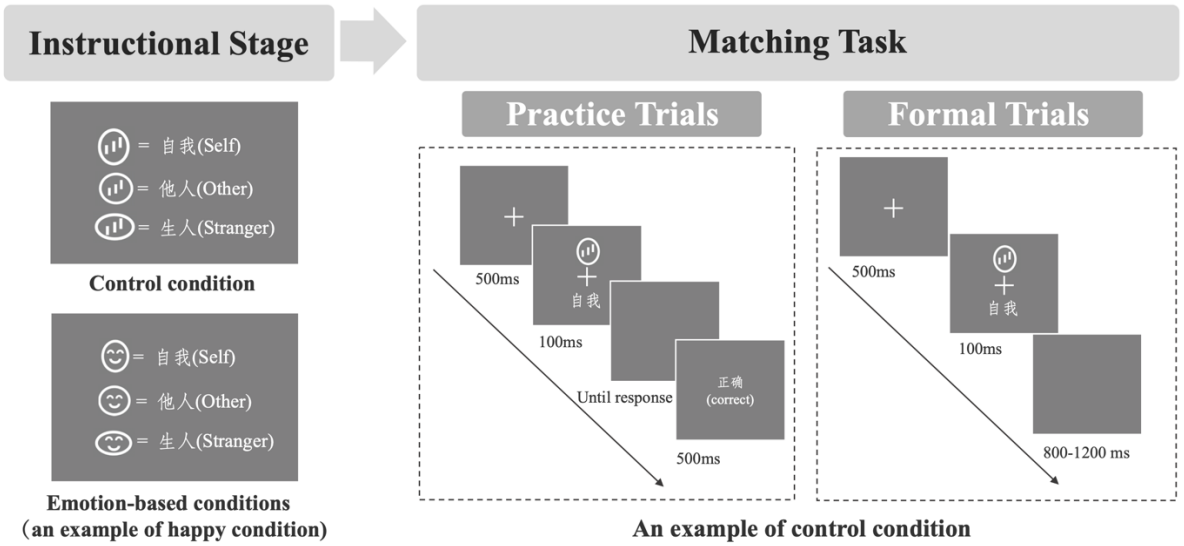

Fig. S1 Procedure of the SMT in Experiment B (Hu et al., 2023). *Note:* The labels and feedback appeared in Chinese in the experiment. In the associative learning task, the matched associations of shapes and labels were counterbalanced between participants. Timely feedback was not provided in formal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

1.2 Paper Selection Procedure

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

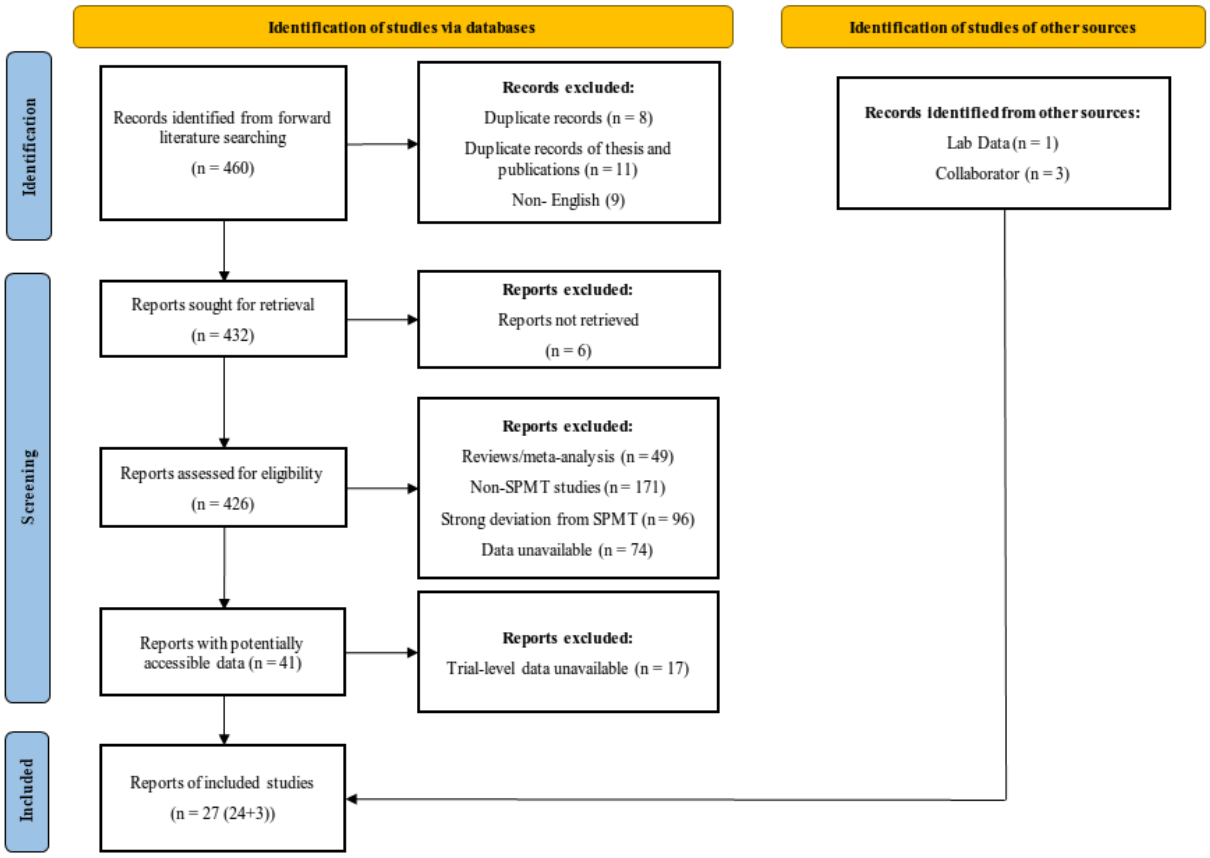

Figure S2. Paper Selection Procedure (adapted from PRISMA Flow Diagram (Page et al., 2021)).

1.3 Parameter Recovery Result for Package Comparison

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

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

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

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

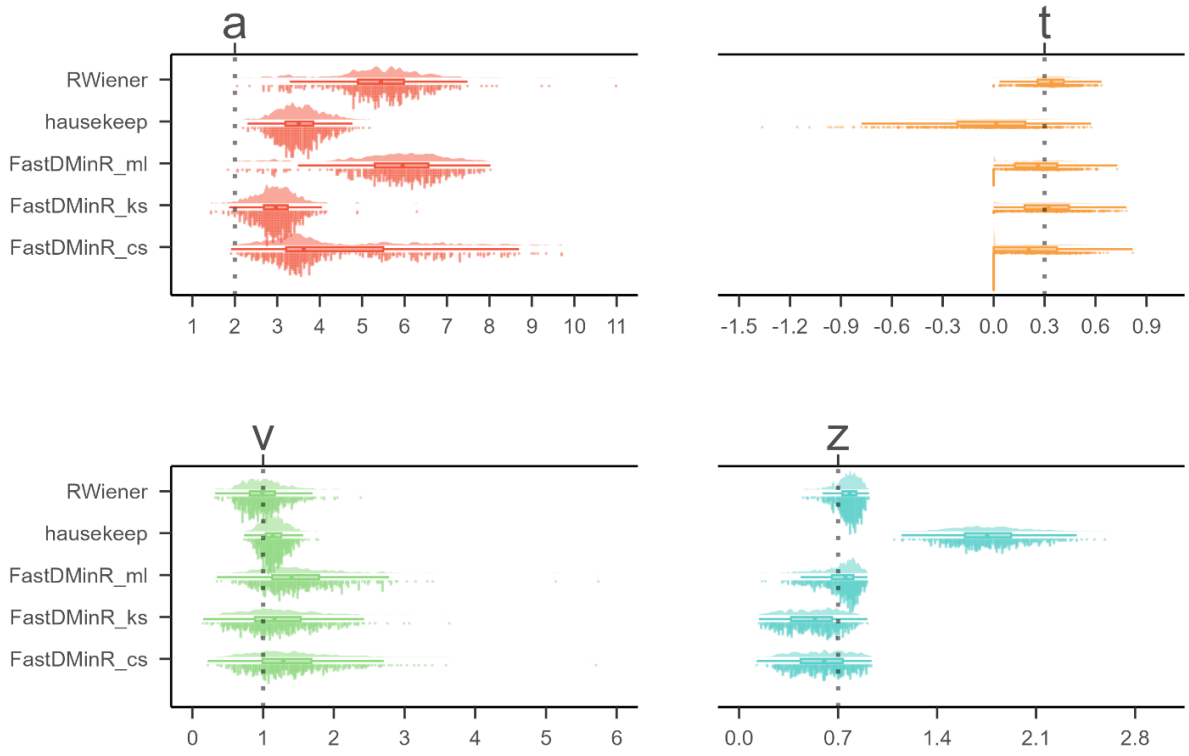

Fig. S3 DDM Packages Comparison. *Note:* The parameters of interest in the Drift-Diffusion Model (DDM) are represented as follows: "*a*" denotes the threshold parameter, "*t*" represents the non-decision time, "*v*" indicates the drift rate, and "*z*" corresponds to the starting point. The y-axis of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

2 Supplementary Results

2.1 Group Level SPE for Other Measures

We conducted a meta-analysis of all the 6 indicators of SPE. The forest plots are presented in Fig. S4.

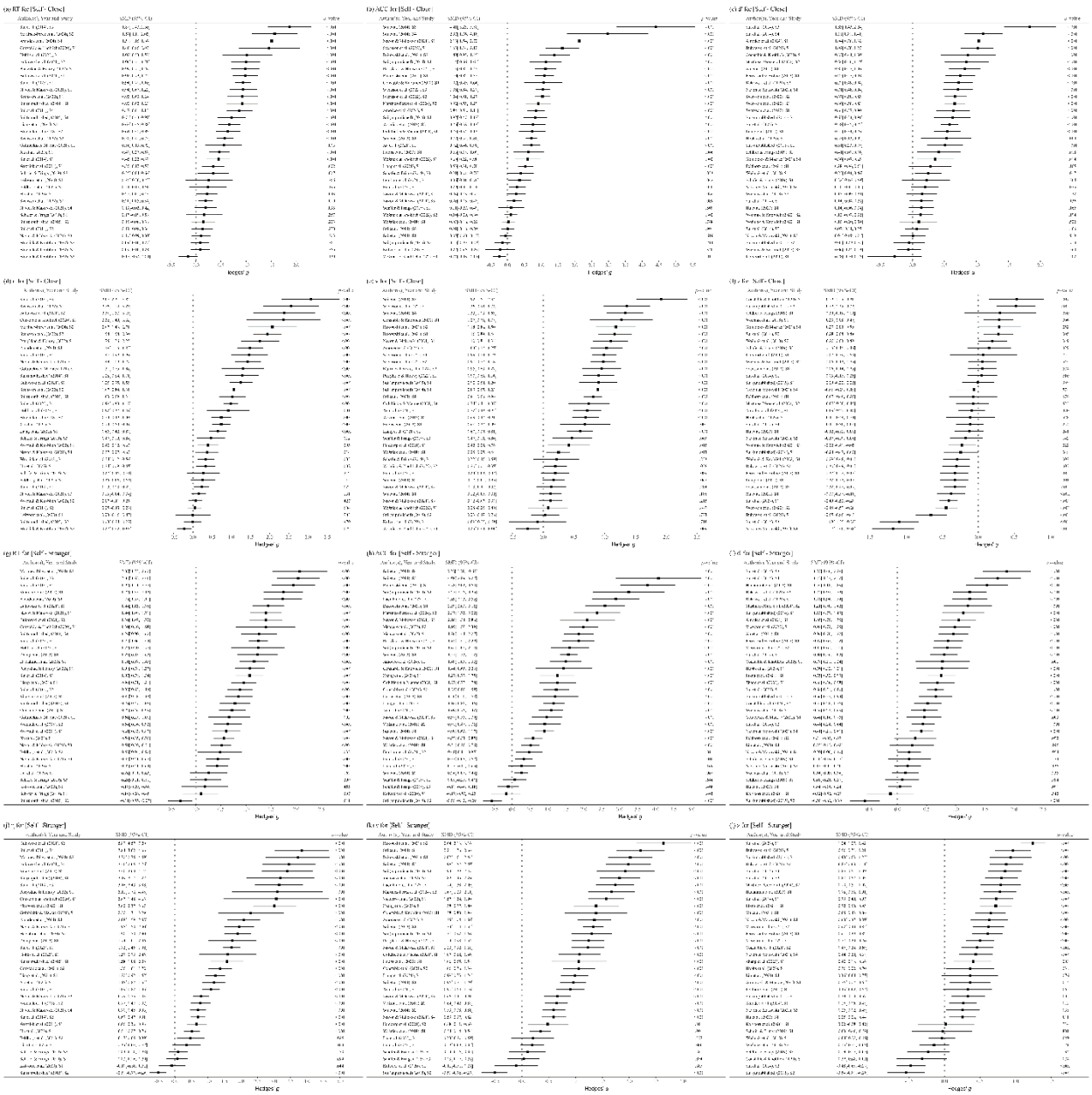

Fig. S4 (a) Forest Plot for SPE Measures. *Note*: Fig (a)-(f) represent the forest plots corresponding to RT, ACC, *d'*, *η*, *v*, and *z* under the condition where Target is Close. Fig (g)-(l) represents the forest plots corresponding to *d'*, *η*, *v*, and *z* under the condition where Target is Stranger.

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

Table S1 T-test Results of SPE Measures in SMT

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

2.2 Split-Half Reliability Using Three Splitting Approaches

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

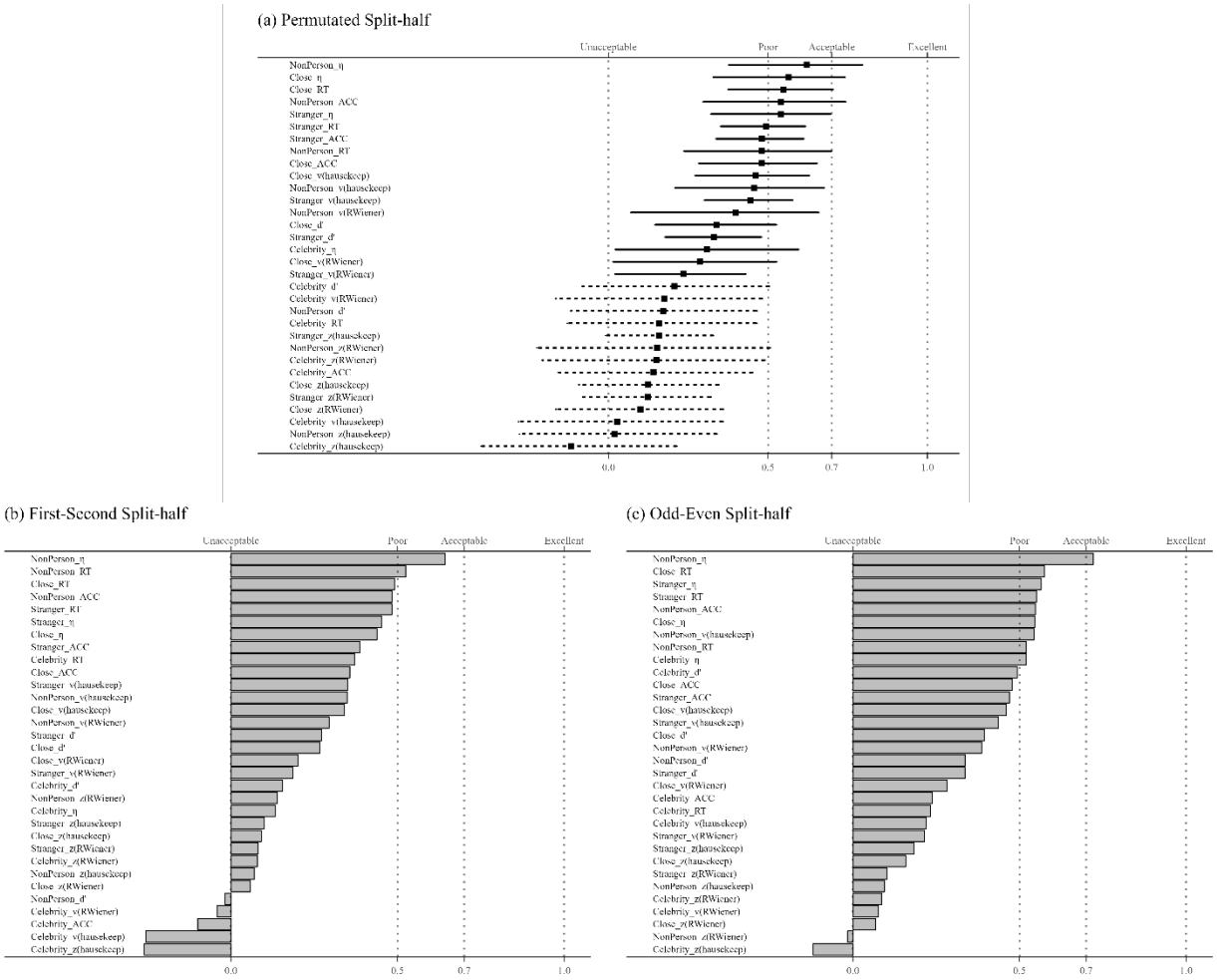

Fig. S5 Results of SHR Using Three Split-half Methods. (a) Results of SHR using Permutated Split-half Methods; (b) Results of SHR using First-Second Split-half Methods; (c) Results of SHR using Odd-Even Split-half Methods. *Note*: The vertical axis of the graph listed 32 different SPE measures, combining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

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

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

2.3 ICCs for SPE Measures Using Another Dataset

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

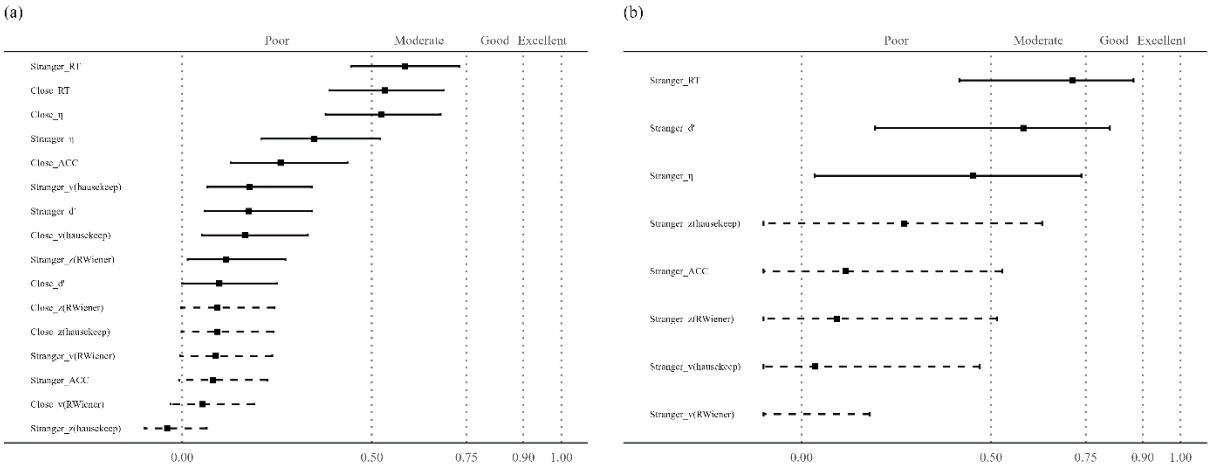

Fig. S6 ICCs for SPE Measures Using Hu et al. (2023) and Another Dataset. (a) ICC2 for SPE measures using Hu et al. (2023); (b) ICC2 for SPE measures using an additional dataset. *Note*: The vertical axis of the graph illustrates eight distinct indicators, which include two additional indices from the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

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

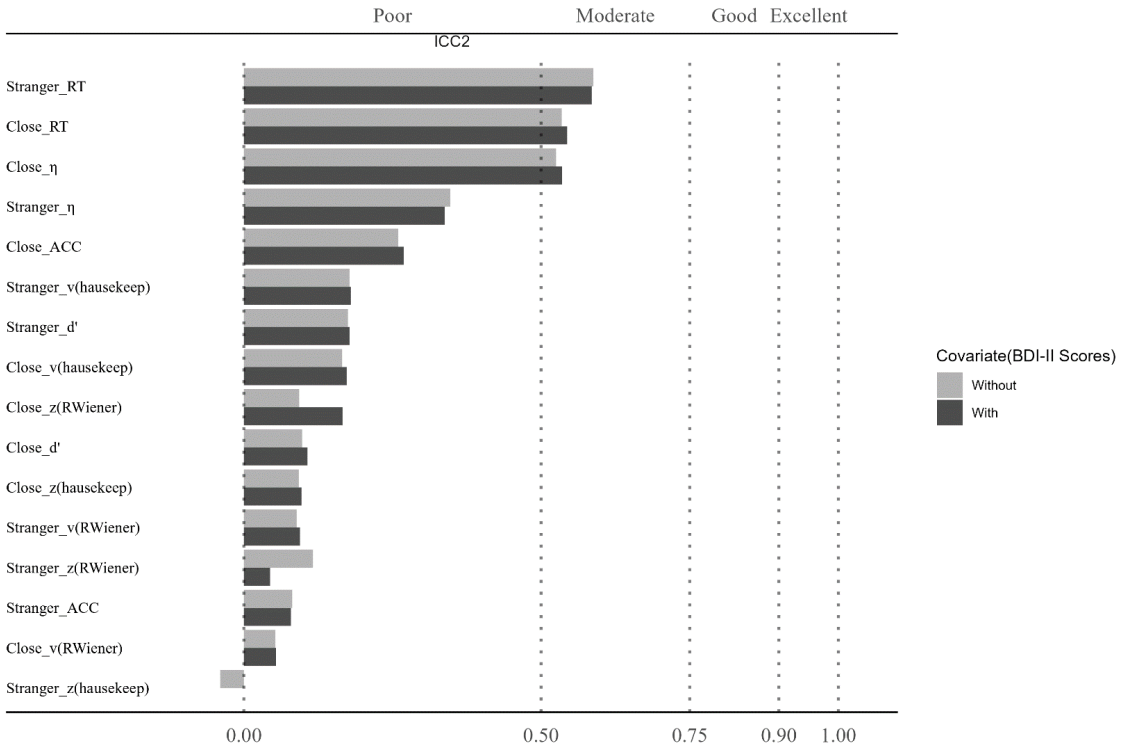

Fig. S7 ICC2 for SPE Measures Using Hu et al. (2023) with Covariant (BDI-II Scores). Note: The vertical axis of the graph illustrates eight distinct indicators, which include two additional indices from the DDM, implemented using the "hausekeep" package. The bar on the graph represents the value of ICC2.

2.4 Exploratory Analysis

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

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

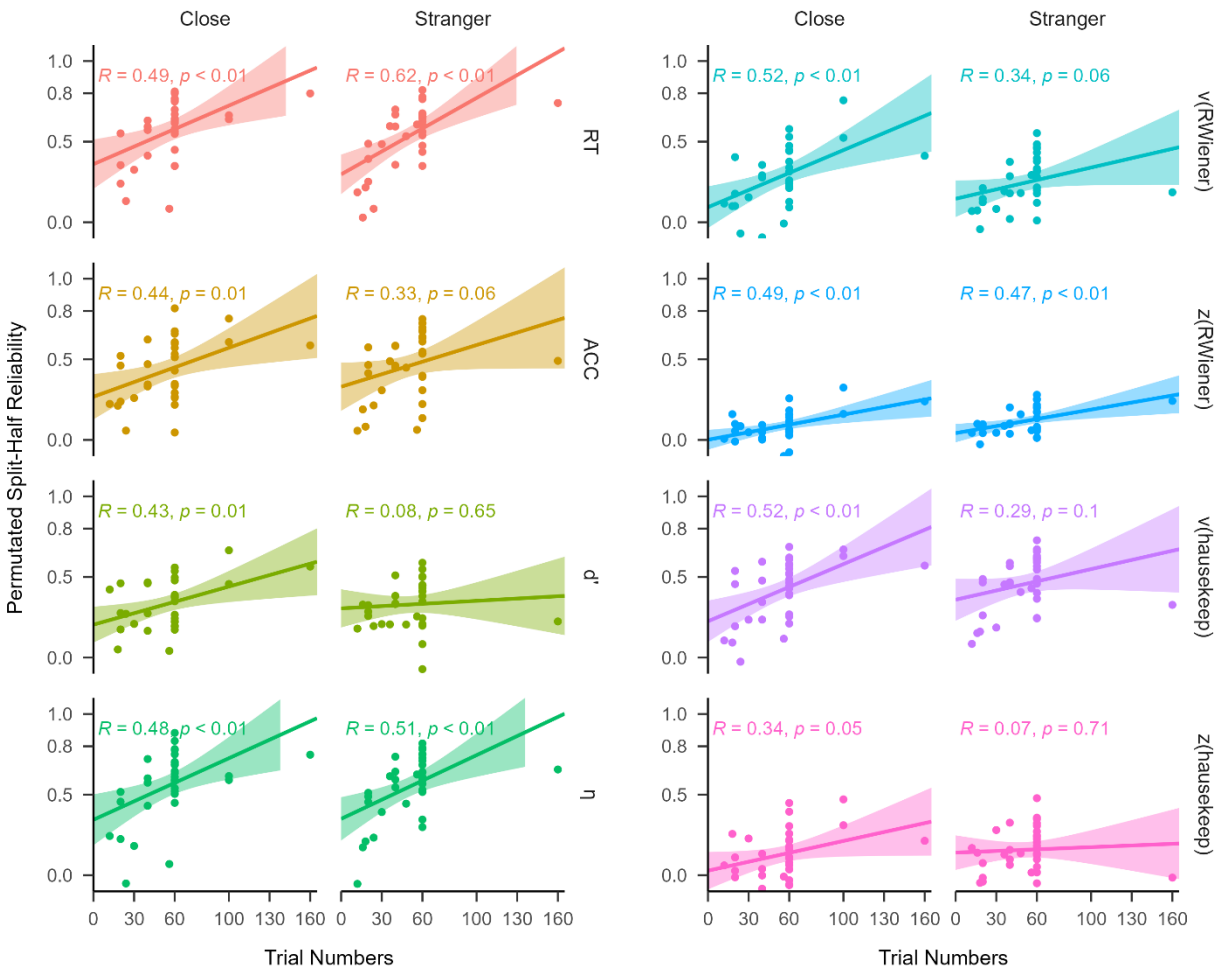

Fig. S8 Regression Analysis Between Permutated SHR and Trial Numbers Using Different SPE Measures. Note: The vertical axis represents the permutated split-half reliability, and the horizontal axis represents the number of trials. Each facet represents one SPE measure.

We also explored the correlation between split-half reliability and effect size (Hedges' *g*) and found mixed results. For some indices of SPE, the correlation between reliability and effect size is significant (e.g., RT, ACC, *d'*, efficiency with stranger as baseline), but for others (e.g., indices with close others as baseline), the correlation was not significant (see Fig. S9). This pattern was consistent with the reliability paradox (Hedge et al., 2018; Logie et al., 1996), suggesting that robust experimental effects are not always associated with robust individual difference correlations.

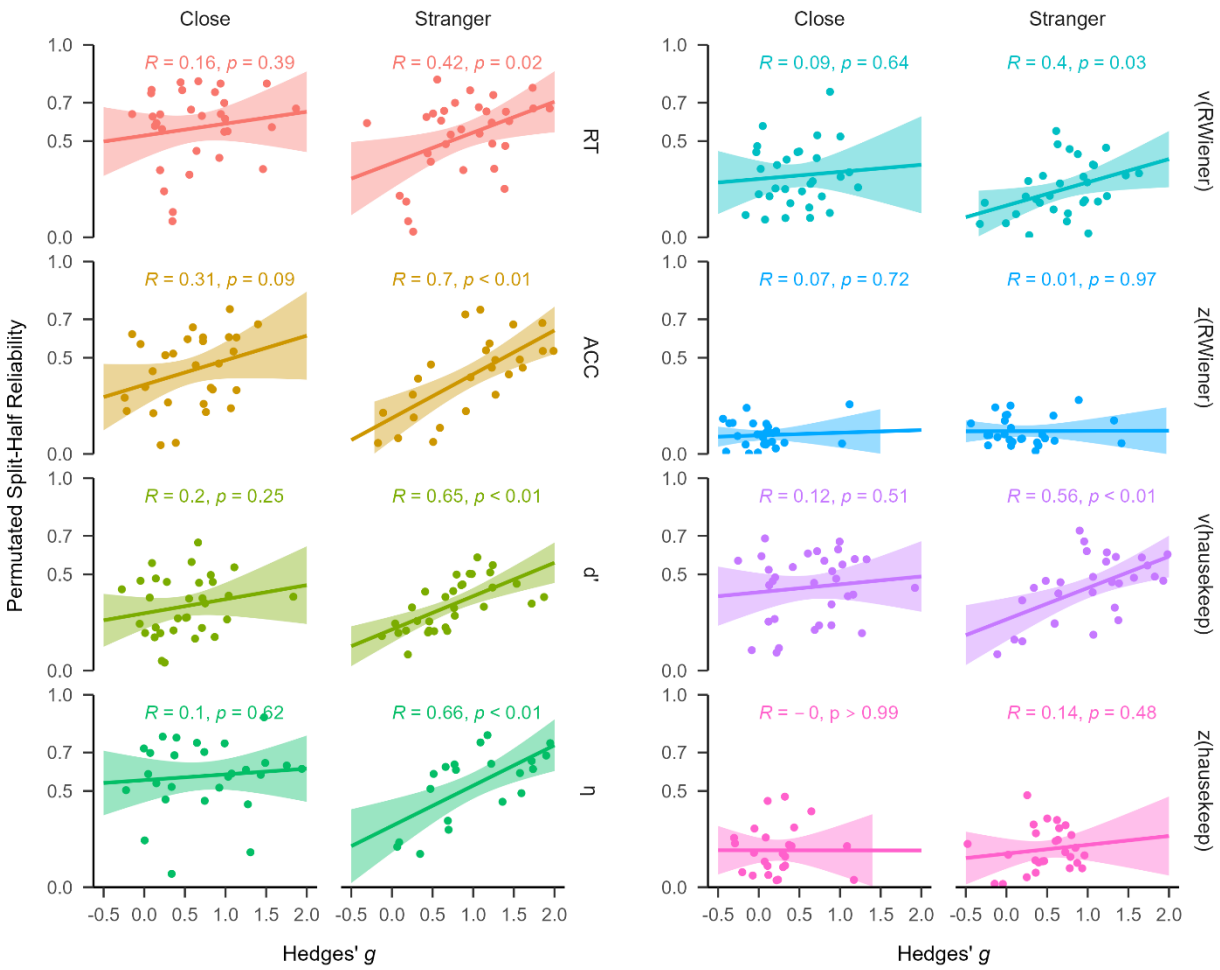

Fig. S9 Regression Analysis Between Permutated SHR and Effect Size (Hedges' g) Using Different SPE Measures. *Note*: The vertical axis represents permutated split-half reliability, and the horizontal axis represents the effect size (Hedges' *g*). Each facet represents one SPE measure.

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

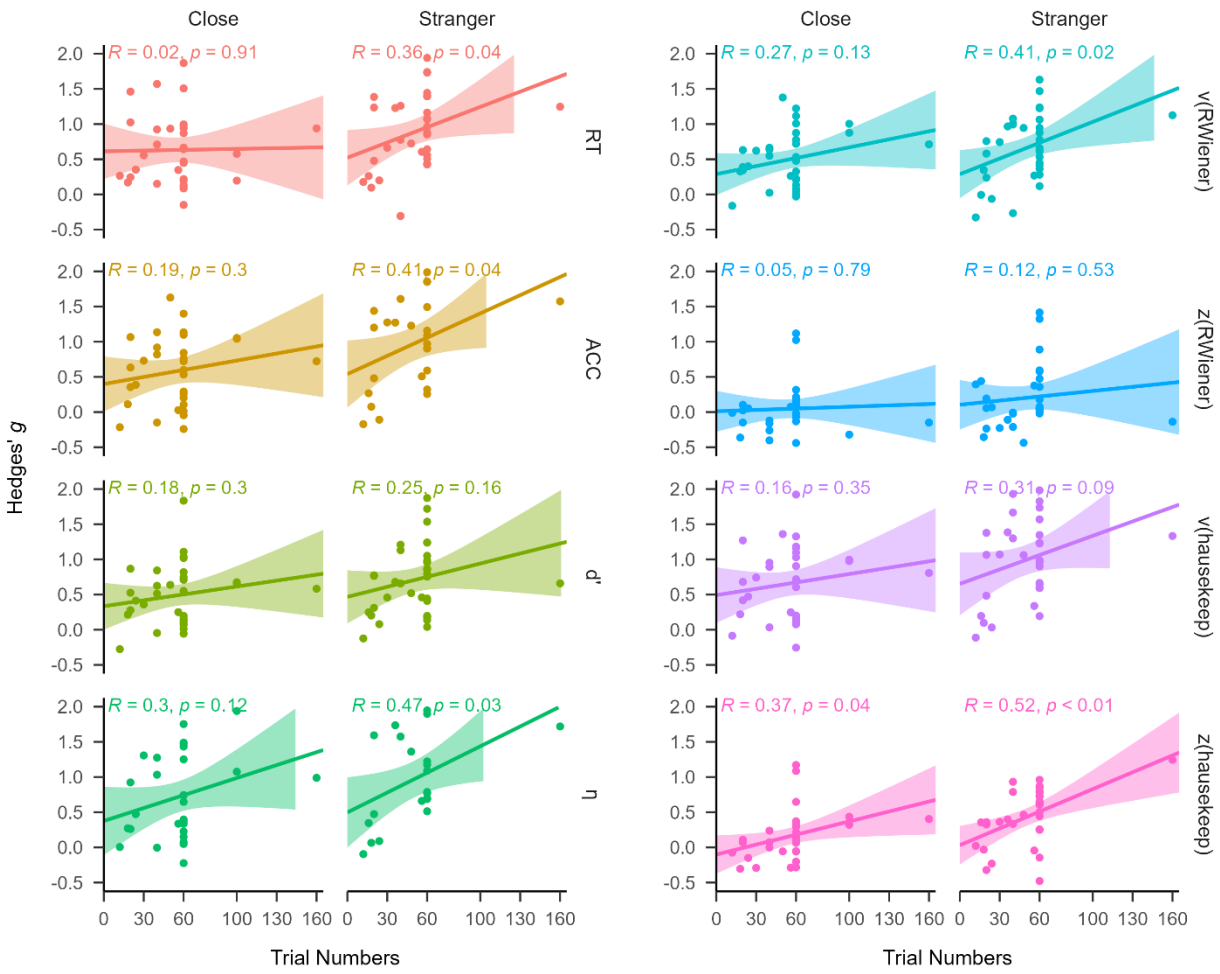

Fig. S10 Regression Analysis Between Trial Numbers and Effect Size (Hedges' g) Using Different SPE Measures. Note: The vertical axis represents the effect size (Hedges' g), and the horizontal axis represents trial numbers. Each facet represents one SPE measure.

Finally, we used the Spearman-Brown prediction formula to predict the trial numbers required for different levels of reliability. The results indicated that the number of trials required for achieving sufficient reliability (e.g., 0.8) varied across different SPE indices. For SPE measured by RT, approximately 180 trials are required to achieve a reliability of 0.8. For other SPE indices, more trials are required (see Fig. S11). It's important to emphasize that these findings are based on our current dataset and should be interpreted with caution. Further validation and verification of this relationship would be essential and will require new data collection efforts in future research.

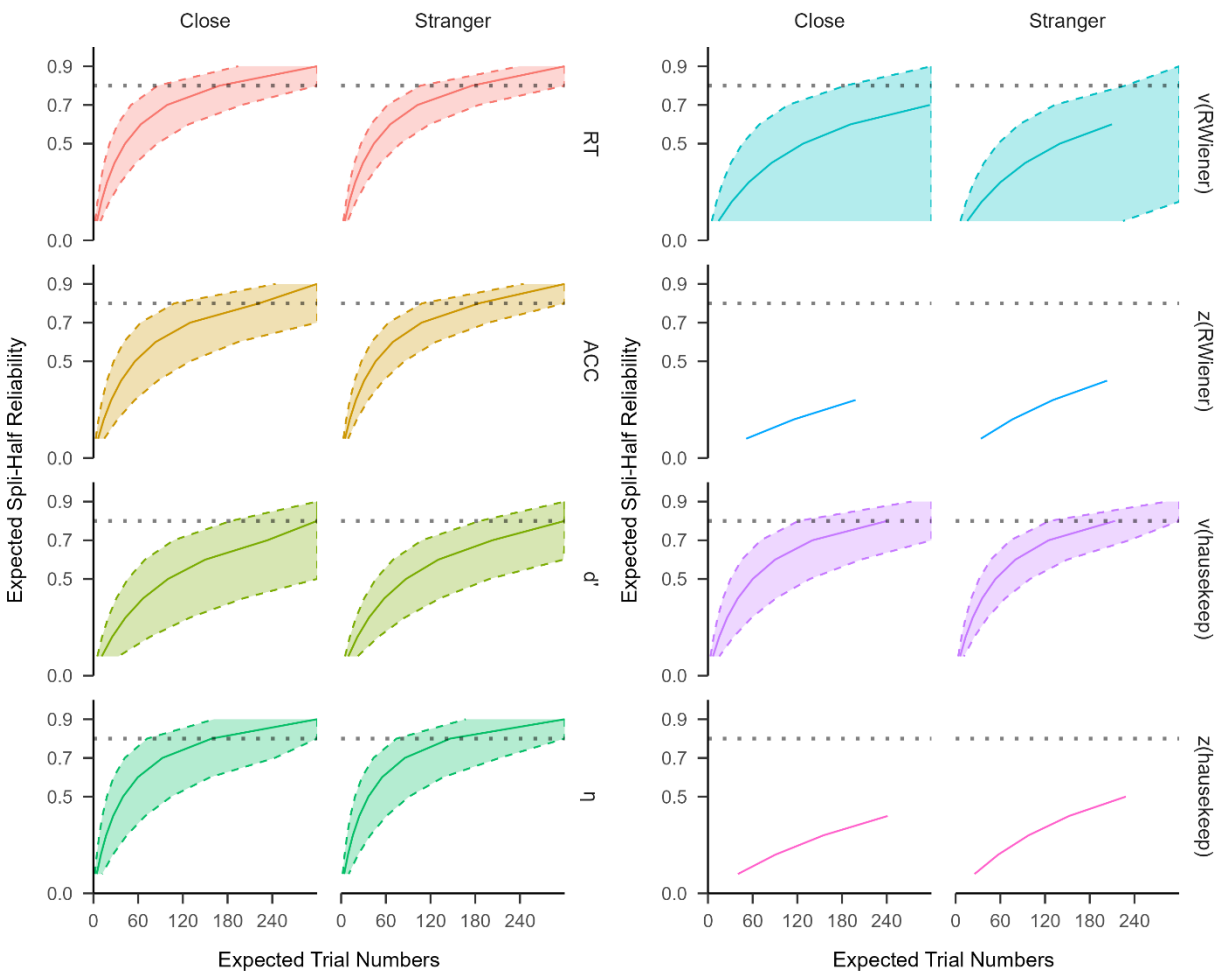

Fig. S11 Expected Trial Numbers Using Different SPE Measures. *Note*: The vertical axis represents the expected trial numbers calculated based on the spearman-brown function, and the horizontal axis represents the expected split-half reliability. Each facet represents one SPE measure. For SPE measured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

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

References

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

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

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

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

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

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

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

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

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

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