nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02198-2

Precise individual measures of inhibitory control

Received: 6 October 2023

Accepted: 1 April 2025

Published online: 28 May 2025

Check for updates

Hyejin J. Lee[1,2,3,8] ✉, Derek M. Smith[4,5,8], Clifford E. Hauenstein[4,6], Ally Dworetsky[1,5], Brian T. Kraus[5], Megan Dorn[5], Derek Evan Nee[1] & Caterina Gratton[1,2,3,5,7] ✉

Inhibitory control is essential to daily function and is a key factor in numerous psychiatric disorders. One popular measure of inhibitory control is the congruency effect, but recent research has highlighted its low reliability, limiting its use for clinical and basic research questions. Here we asked whether it is possible to obtain precise individual estimates of the congruency effect. We sampled more than 5,000 trials from nine participants across four inhibitory control tasks. This dataset, made public for the community, demonstrates that precise individual estimates are achievable but with higher numbers of trials than typically collected with common tools. Using a combination of datasets and simulations, we show that extensive sampling is necessary to reveal true individual differences and improve observations from alternative modelling approaches. We share our dataset as a resource to further understand sources of variation in inhibitory control, ultimately advancing research in this critical field.

Inhibitory control refers to the ability to resist interference and suppress dominant responses to carry out goal-directed behaviour[1–8]. Inhibitory control helps to enable everyday activities and achieve goals on diverse timescales[9–11]. Deficits in inhibitory control have been implicated in a number of psychiatric disorders, including obsessive–compulsive disorder[12,13], schizophrenia[14,15] and attention deficit hyperactivity disorder[16]. Accordingly, multiple efforts are underway to understand variation in inhibitory control across individuals as well as within a person over the lifespan, and to connect this variation with neurobiological markers. These include large consortium data collection projects, such as the Adolescent Brain Cognitive Development (ABCD) study[17].

Inhibitory control in laboratory settings is often examined with tasks that contrast the impact of high-conflict distractors (that is, incongruent trials) relative to low conflict situations (congruent trials). The difference between these types of trials is referred to as the congruency effect[18–20] and is taken to be a reflection of control (note that some may suggest that 'interference control' is more appropriate for describing this process with a mechanistic focus; however, given that 'inhibitory control' is more commonly used (for example, National Institutes of Health (NIH) Toolbox Flanker Inhibitory Control and Attention Test), we also use it here without committing to any specific underlying mechanisms of control). The congruency effect has been extensively tested and highly replicated[21]. It has also been suggested that the congruency effect is stable across time[22], indicating that it may represent a trait-like characteristic of a person. These properties have prompted many to use the congruency effect to investigate both individual-level inhibitory control within and between individuals[23–25] and relate this variation to neural activity[3,26–29].

However, poor reliability is likely to be a major obstacle in these enterprises[30]. Several studies have reported low reliability in measures of inhibitory control, particularly in the widely used congruency effect,

[1]Department of Psychology, Florida State University, Tallahassee, FL, USA. [2]Department of Psychology, University of Illinois Urbana-Champaign, Champaign, IL, USA. [3]Beckman Institute, University of Illinois Urbana-Champaign, Champaign, IL, USA. [4]Department of Neurology, Division of Cognitive Neurology/Neuropsychology, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [5]Department of Psychology, Northwestern University, Evanston, IL, USA. [6]School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA. [7]Neuroscience Program, University of Illinois Urbana-Champaign, Champaign, IL, USA. [8]These authors contributed equally: Hyejin J. Lee, Derek M. Smith. ✉e-mail: leex6248@gmail.com; cgratton@illinois.edu

https://doi.org/10.1038/s41562-025-02198-2

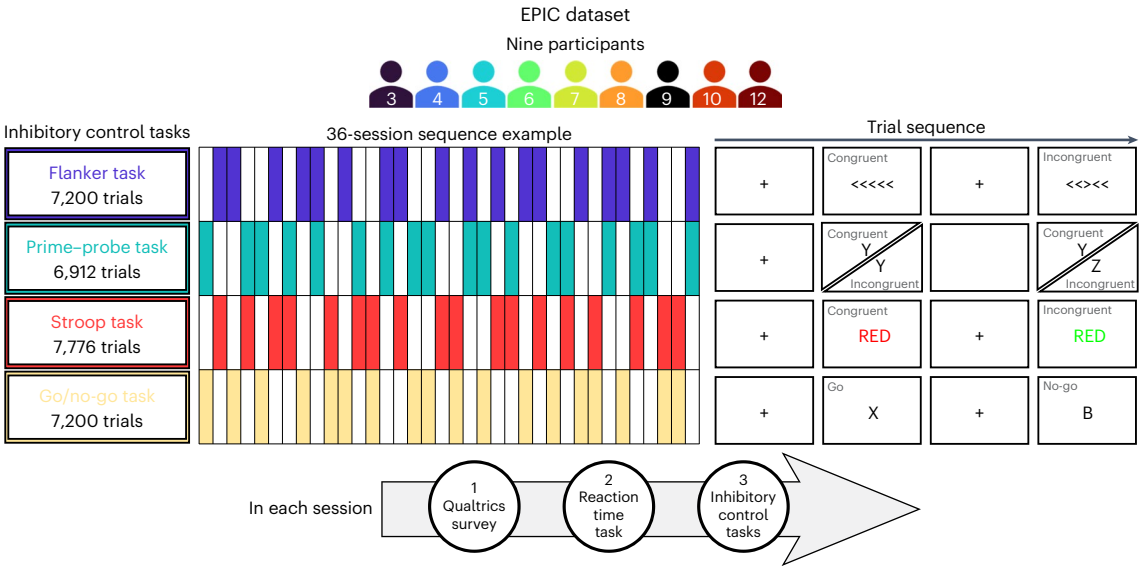


Fig. 1 | EPIC dataset. Nine participants (5 females and 4 males, ages 18–30 years; one participant was excluded from the current analyses) completed 4 inhibitory control tasks across 36 sessions (with 2 tasks per session; the order was counterbalanced across participants). Exemplar trial sequences of the tasks are shown. Modifications to the task stimuli and background colour were made for visualization purposes. Each session started with a Qualtrics survey assessing daily activities and mood, followed by a simple reaction time task measuring reaction time to a single visual stimulus (that is, a white square). Two of the four inhibitory control tasks were then administered. We also collected Big Five personality traits. Participants were tested either in the laboratory or at home. More details on the dataset and experimental designs can be found in the Methods. The dataset is available at https://osf.io/jk9nb/.

which relies on reaction time difference scores[31–33]. This issue limits the ability of these measures to be used for clinical applications[34,35] and for predicting self-regulation in the real world[36,37]. Low reliability may also confound theoretical interpretations to other control functions[33,38,39]. Importantly, poor reliability is a major obstacle to identifying relationships between the brain and behaviour[40]; recent studies have demonstrated that, when the reliability of behavioural measures is low, brain-based prediction of the behavioural phenotype is extremely poor[41,42]. Unsurprisingly, given this background, inhibitory control measured with the NIH Toolbox Flanker Inhibitory Control and Attention Test had one of the lowest brain-based prediction accuracies in the Human Connectome Project (HCP) dataset[43].

The low reliability of the congruency effect has been attributed to a combination of high measurement error (leading to low within-participant precision) and limitations in estimating between-participant variation[30,32,39]. Some have suggested that estimates of the congruency effect could be improved by recruiting more diverse samples[44], modifying experimental designs[45] or using modelling approaches, such as drift–diffusion modelling[36,46], factor analysis[47] or hierarchical modelling[33,39,48]. However, to reduce measurement error, the most straightforward approach is to increase the number of trials collected per participant. Past research has used simulations to call attention to the need for larger trial numbers to increase reliability[39]. As of yet, it remains an open question whether it is possible to achieve a sufficient level of precision in individual-level estimates and what trial number is needed for these estimates to have utility as a phenotypic marker.

Here, we sought to empirically estimate the peak precision possible for the congruency effect once measurement error was reduced by collecting a very large number of trials. To this end, we collected data from 9 participants across 36 sessions as they completed 4 inhibitory control tasks (3 of which included congruency effects). We call the dataset EPIC or the Extended Precision measurement of Inhibitory Control. Using this approach, we sought to determine the maximal precision of the congruency effect that can be achieved within individuals. We used this approach to ask whether increasing trial numbers effectively reduces measurement error without introducing systematic variability related to repeated testing. We also investigated how many trials are needed to achieve the desired levels of reliability and whether this number is feasible without requiring prohibitively long testing. Finally, using this dataset, along with other public datasets and simulations, we examined how collecting more trials affects estimates of between-participant variability and the performance of advance modelling approaches. We provide the EPIC dataset as a resource for the community to further assess measurement properties of inhibitory control within individuals over time and as bases for simulation studies and model validation to benchmark new analysis methods. These investigations are essential to determine whether inhibitory control measures are useful to pursue in basic and clinical research projects.

Results

Overview

To examine the precision of inhibitory control measures, we collected EPIC, a dataset with extensive sampling from 9 participants who were tested on 4 inhibitory control tasks across 36 sessions (Fig. 1). This dataset includes 3 tasks with congruency effects: a flanker task (with a total of 7,200 trials per participant), a prime–probe task (6,912 trials per participant) and a Stroop task (7,776 trials per participant). The dataset also includes a fourth inhibitory control task (go/no-go, 7,200 trials per participant) that was not analysed here as we focused on examining the congruency effect. The data from all four tasks are made into a publicly available resource associated with this publication.

In addition, we replicated and extended the findings from the EPIC dataset using two publicly available datasets: (1) a dataset by Robinson and Steyvers[49], collected online through Lumosity, consisting of 495 participants in a flanker task with 491–5,939 trials per participant, and (2) a dataset by Hedge et al.[32], collected in-person, consisting of 112 participants in the flanker and Stroop tasks with 1,440 trials per participant. To aid in interpretating these results, we also generated simulated models based on all three datasets.

The results are organized into three major groups. The first group of results ('Congruency effects can be measured with high precision' to 'Precise congruency effects need more than 1,000 trials' sections) shows empirical results from our EPIC data to assess the efficacy of extensive testing. We examined (1) peak individual-level precision

https://doi.org/10.1038/s41562-025-02198-2

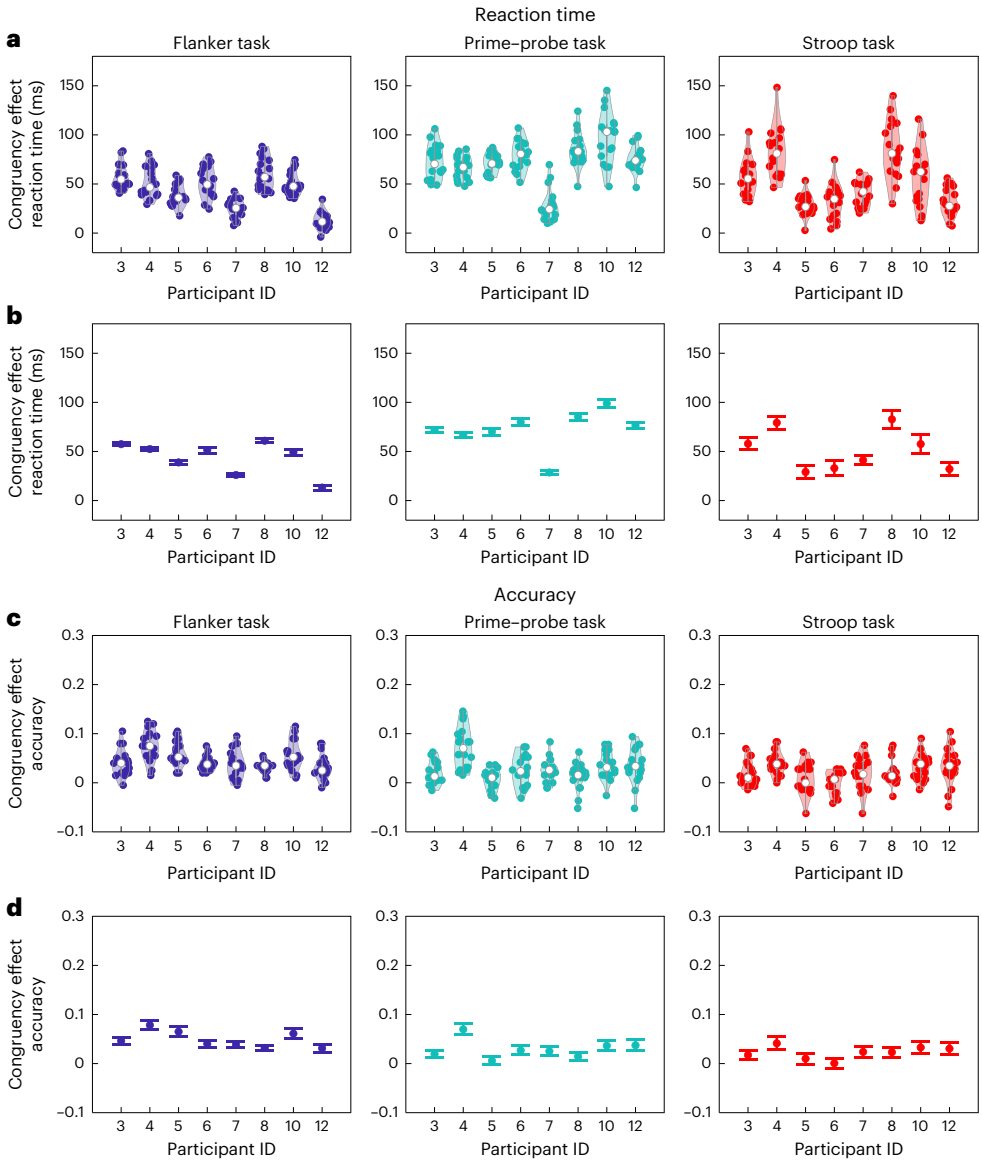


Fig. 2 | Large within-participant variability versus precise individual estimates. a, The session mean reaction time for the congruency effect. Each violin plot represents a participant's mean congruency effect across sessions (one dot per session, totalling 18). The plot depicts the probability density of the congruency effect across trials (400 trials for the flanker task, 384 for the prime–probe task and 288 for the Stroop task; note that the total for the Stroop task excludes neutral trials). The white dot indicates the median of the 18 sessions. The height of the density plot, which reflects the relative frequency of congruency effect values, shows an extended oblong shape rather than clustering around a single value, suggesting large variability within participants across sessions. **b**, Grand mean reaction time across all sessions (7,200 trials for the flanker task, 6,912 for the prime–probe task and 5,184 for the Stroop task). The error bars represent the 95% confidence intervals of the mean, calculated from 1,000 bootstraps. These error bars indicate that highly precise individual estimates of congruency effects are possible when sampling more than 5,000 trials. **c**, The session mean accuracy for the congruency effect, illustrated in violin plots. Consistent with **a**, large session-level variability within participants is observed. **d**, The grand mean accuracy across all sessions and its error bars. Consistent with **b**, the accuracy data demonstrate substantially smaller variability with more than 5,000 trials compared with 400 trials. Performance was generally high with a smaller congruency effect on accuracy than on reaction time across participants. Accuracy and reaction time for each condition (congruent or incongruent) are shown separately in Supplementary Figs. 2 and 3. IES are presented in Supplementary Figs. 1 and 4.

in the congruency effect, (2) temporal effects related to repeated testing and (3) the number of trials needed to obtain reliable congruency effects. The second group of results ('Repeated measures reduce variability in congruency estimates' to 'Trial counts matter more than participants in reliability' sections) presents empirical and simulation results from the two public datasets with larger samples to replicate and extend our EPIC results. We examined (1) the impact of trial numbers on within- and between-participant variability, (2) how extended trial sampling ensures stable results for both estimates, (3) the impact of within-participant variability on between-participant variability and (4) how increasing the number of trials and participants affects reliability. The final section ('Advanced models benefit from more within-participant data' section) uses all three datasets in combination with empirical and simulation results to demonstrate how trial numbers influence diverse modelling approaches (drift–diffusion modelling, factor analysis and Bayesian hierarchical modelling) used to address reliability concerns in inhibitory control.

Congruency effects can be measured with high precision

First, we examined the maximal precision with which congruency effects can be measured in each participant once sampling variability is addressed using our extensively sampled dataset.

Figure 2 shows each participant's mean reaction time and accuracy for the congruency effect. A comparison of the violin plot for 18 sessions with the grand mean of all sessions shows that precise individual-level congruency effects can be achieved with sufficient sampling. Across participants, the average 95% confidence interval for the reaction time congruency effect in any given session (288–400 trials) is 17.98 ms for the flanker task, 25.33 ms for the prime–probe task and 58.45 ms for the Stroop task. Considering the grand mean congruency effects (44.01 ms for flanker, 73.98 ms for prime–probe and 58.20 ms for Stroop) along with the between-participant standard deviation (15.51 ms for flanker, 19.70 ms for prime–probe and 28.05 ms for Stroop), these are high levels of error. Consider randomly selecting a single session point from each participant: with this level of variation, rank ordering participants would frequently be inaccurate.

By contrast, the congruency effect calculated from the full 5,184–7,200 trials across all sessions indicates that single individuals can have precise congruency effect estimates. With this number of trials, the 95% confidence intervals are, on average, 4.66 ms for the flanker task, 6.33 ms for the prime–probe task and 14.12 ms for the Stroop task. These values are much smaller than those observed for a single session's worth of data. These estimates are precise enough to show consistent interparticipant differences even in this small sample of participants. For example, EPIC 07 shows relatively small congruency effects in all three tasks whereas EPIC 08 shows large congruency effects. The rank orderings between EPIC 05 and 06, 07 and 08, and 10 and 12 are consistent across the tasks. However, some effects differ by task. For example, the relative positions of EPIC 03 and 04 swap between the flanker and Stroop tasks (see Extended Data Fig. 1 for rank order consistency across the three tasks).

Accuracy results also demonstrate that more precise estimates can be achieved with larger amounts of data. We also found consistent results in the inverse efficiency scores (IES), which combine reaction time and accuracy to account for speed–accuracy trade-offs (Supplementary Fig. 1). For a summary of the individual grand means and standard errors of reaction time, accuracy and IES, see Supplementary Table 1.

In addition, we replicate previous findings[31], showing that session-by-session variability is lower for incongruent and congruent performance when measured separately (Supplementary Figs. 2–4). This is because difference scores are associated with an increase in sampling variability; when the two components of a difference score are correlated, as is the case with congruent and incongruent trials, the subtraction removes reliable variance, increasing the proportion of variance attributable to error[30,31,50,51]. Given recent suggestions to replace the congruency effect with incongruent trial performance[31,52], we conducted an analysis on the rank order consistency between congruency effects and incongruent trial performance to examine whether the two are measuring similar constructs using the two public datasets (Extended Data Fig. 2). The results demonstrate that they are correlated, but inconsistency in rank orders also exists, lending caution to the idea of substituting incongruent trial performance alone for the congruency effect.

Extensive data collection is feasible

While extensive repetition reduces session-level variability within individuals, it may introduce variability related to temporal effects. We examined three potential effects in our dataset to evaluate the practicality of collecting large numbers of trials: (1) performance improvement over time, (2) performance impairment across sessions and (3) performance impairment within sessions. Performance improvement may be linked to practice or learning effects over time. Performance impairments, whether across or within sessions, could be due to a variety of sources, such as participants losing interest, motivation, concentration or experiencing increased fatigue. Note that our participants performed 2 sessions for each task per week for 9 weeks. Single sessions included 2 tasks and lasted for ~45–60 min.

Extended Data Fig. 3 shows decreases in the magnitude of the congruency effect with additional experience in the tasks. These effects are clearest in reaction time but vary across participants and tasks. These effects are absent in the dataset of Robinson and Steyvers[49], which also acquired hundreds to thousands of trials but across years. Thus, variability in temporal sampling may play a crucial role in observing these improvement effects. Notably, across all sessions and tasks, all participants retained congruency effects in their reaction time measures despite the presence of these improvement effects.

Extended Data Figs. 4 and 5 show that no strong signs of performance deterioration across and within sessions are observed in our dataset, although a small level of performance degradation within sessions was observed in the Stroop task. Together, these findings suggest that collecting extensive amounts can be feasible under certain conditions, which we will elaborate on in the Discussion.

Precise congruency effects need more than 1,000 trials

Next, we asked how many trials are needed to get precise estimates of the congruency effect within individuals. We assessed the precision of individual-level measures by systematically increasing the number of trials and identifying the point at which estimates of within-participant variability in the congruency effect begin to stabilize near zero. This stabilization point approximates the point at which additional trials yield diminishing returns relative to the effort required to collect them. Note that the exact level of precision needed will depend on particular questions and applications. Although we focus on recommendations based on the largest improvements in reliability, we provide the full stability curves in the Article for readers interested in other applications.

We utilized two methods that assessed the within-participant variability, as our focus was on achieving precise individual-level measures. As we will demonstrate in the next section, this stabilization point based on within-participant variability consistently aligns with stabilization estimates derived from between-participant variability.

In our primary method, we randomly split in half each participant's data into two sets, a reference set and a test set (trial data were split in contiguous segments; Methods). The reference set was used as our best estimate score based on a large amount of independent data (~3,000 trials). We then extracted increasingly large subsets of data from the test set. We compared congruency effects between the accumulating test set samples and the reference set, and the stabilization point was established as the approximate location where the absolute difference flattens out (this may not ever reach zero, as some level of error due to, for example, state-based variability, may remain). This approach provides an estimate of the precision of the congruency effect for each individual by comparing it with the best measure available. Note that before running this analysis, we removed the improvement effects (decreasing congruency effects over time; Extended Data Fig. 3) using linear regression (see Extended Data Fig. 6 for before-and-after linear regression; results were similar even when the improvement effects were retained in the data as shown in Supplementary Fig. 5).

Our results show that more than 1,000 trials are needed for the test samples to be comparable to the reference set (Fig. 3). At 1,000 trials, the test samples show an average absolute difference of 4.04 ms per individual for flanker (9.18% of the grand mean of 44 ms), 5.43 ms for prime–probe (7.43% of the grand mean of 74 ms) and 9 ms for Stroop task (15.52% of the grand mean of 58 ms). Although smaller, additional gains in precision are seen beyond 1,000 trials as well.

These results are replicable using our second method for assessing within-participant reliability, which calculates the width of the 95% confidence interval for the mean congruency effect with bootstrapping at varying numbers of trials (Methods and Supplementary Fig. 5). The width of the confidence interval reflects within-participant variability in the congruency effect, and the stabilization point—where the

https://doi.org/10.1038/s41562-025-02198-2

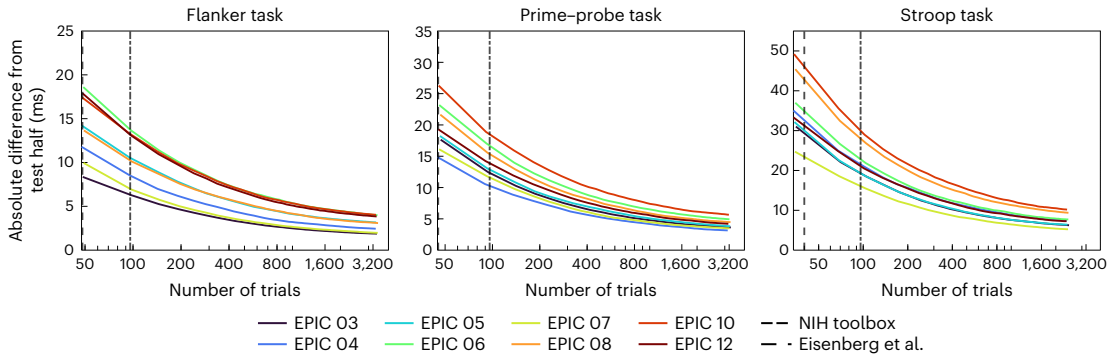


Fig. 3 | Precision of congruency effect estimates with different trial numbers. This figure shows the number of trials required to get stable congruency effect estimates, shown as reaction time stability curves for each participant across the three tasks. The absolute difference between the reference set and the accumulating test set sample is plotted as a function of number of trials (on a logarithmic scale). We defined the stabilization point as the point where the difference appears to level off, indicating that additional trials provide relatively smaller benefits (see 'Within-participant precision of the congruency effect' section in the Methods for details on the analysis). We estimate this number as 1,000 trials. The curves are comparable across participants and tasks, but the precision is highest for the flanker task and the lowest for the Stroop task, given the same number of trials. This suggests that, while more sampling improves precision until at least 1,000 trials, the extent to which precision is improved will be affected by particulars of the dataset and the individual participant. The two vertical lines mark the number of trials collected in the NIH Toolbox Flanker Inhibitory Control and Attention test (dashed line at 40 trials) and Eisenberg et al.'s[36] Stroop task (dashed–dotted line at 96 trials). We also plotted the absolute difference as a percentage of an individual's grand mean to illustrate how small the within-participant variability can be (for example, 20% of the grand mean; Supplementary Fig. 6). In addition, similar stability curves for accuracy are shown in Supplementary Fig. 7, and for IES in Supplementary Fig. 8. Our Stroop task includes a neutral condition, allowing us to separate congruency effects for the two conditions (facilitation: neutral-congruent; interference: incongruent-neutral); the stability curves for the facilitation effect and the interference effect are shown in Supplementary Fig. 9. Finally, see Supplementary Fig. 10, which plots the correlation between the reference set and the test set across participants.

width levels off—approximates the true within-participant variability. Correlation coefficients between the data from the two methods for each participant across each task are $r \geq 0.98$, suggesting highly consistent results.

The number of trials needed to achieve these relatively precise congruency effects (>1,000) exceeds what is typically collected in most standard study designs. In experimental research, which compares group averages of a sample size of about 30, it is typical for 500–800 trials to be administered per participant[53–55]. For cross-participant correlational research, the required sample size is often much larger (that is, hundreds of participants), which typically comes at the cost of the number of trials per participant[42,56]. For example, 96 trials were collected per person in Eisenberg et al.'s[36] Stroop task, and 40 trials in the NIH Toolbox Flanker Inhibitory Control and Attention Test[57].

With 96 trials, the 95% confidence interval, averaged across participants, is 60 ms in the flanker congruency effect (see dashed–dotted lines in Supplementary Fig. 5). With 40 trials, the 95% confidence interval is 87 ms (dashed lines). These levels of error are substantial, given that the average flanker congruency effect is 44 ms with a between-participant standard deviation of approximately 17 ms. Even larger errors are seen with these trial numbers in the prime–probe and Stroop tasks.

Repeated measures reduce variability in congruency estimates

We have shown that repeated measures can successfully reduce within-participant error. The next important question is how repeated measures affect between-participant variability. If testing more trials continuously reduces between-participant variability, making measures highly comparable, extensive testing would have limited benefits for improving the reliability of the congruency effect. Alternatively, if it helps to reveal stable between-participant variability once within-participant error has been minimized, then achieving high levels of reliability would be possible with extensive testing.

To test these predictions, we turned to a publicly available dataset from Robinson and Steyvers[49], which includes online flanker task data from 495 participants. The number of trials per participant ranges from 491 to 5,939 trials. For the purposes of this study, we limited analyses to participants with good accuracy rates (>70% on average, no sessions with 0% accuracy) and more than 2,500 correctly responded trials (Methods). This left a total of 185 participants for analysis.

First, we replicated the findings from the EPIC dataset, showing that in this larger sample, within-participant variability decreases with greater numbers of trials, plateauing around 1,000 trials (Fig. 4a,c). This shows that the EPIC results are robust, even when tested in a larger and more heterogeneous group.

Importantly, as the within-participant error decreases, so does the between-participant standard deviation (Fig. 4b,d). Consistent with the trajectory of within-participant variability, between-participant standard deviation stabilizes after acquiring about 1,000 trials. These findings extend our estimation of the approximate number of trials needed to achieve a high level of individual precision in the assessment of reliable individual differences.

High trial sampling stabilizes between-participant variability

The association between within- and between-participant variability may be driven by large within-participant error, which can confound between-participant variability estimates[33,39,58,59], as expected from statistical analysis (Supplementary Equation 1). This contamination arises because we split the analysis of multilevel data into two steps when calculating the congruency effect: we first calculate the mean reaction time for each participant and then calculate the difference in these mean reaction time for congruent and incongruent trials. Critically, in the second step, we treat these measures as fixed and known, without systematically accounting for the imprecision in their estimates. As this imprecision is unaccounted for, between-participant variability is contaminated by within-participant variability (see also refs. 33,58).

To provide improved intuition for this relationship, we created two simulated models with cases of small versus large trial sampling (Fig. 5). With small trial sampling, between-participant standard deviation would decrease substantially as trial numbers increase, due to the high influence of within-participant variability. In contrast, when trial numbers are high (and, therefore, within-participant variability is relatively low), we hypothesized that between-participant standard deviation would remain stable, revealing its true estimate.

https://doi.org/10.1038/s41562-025-02198-2

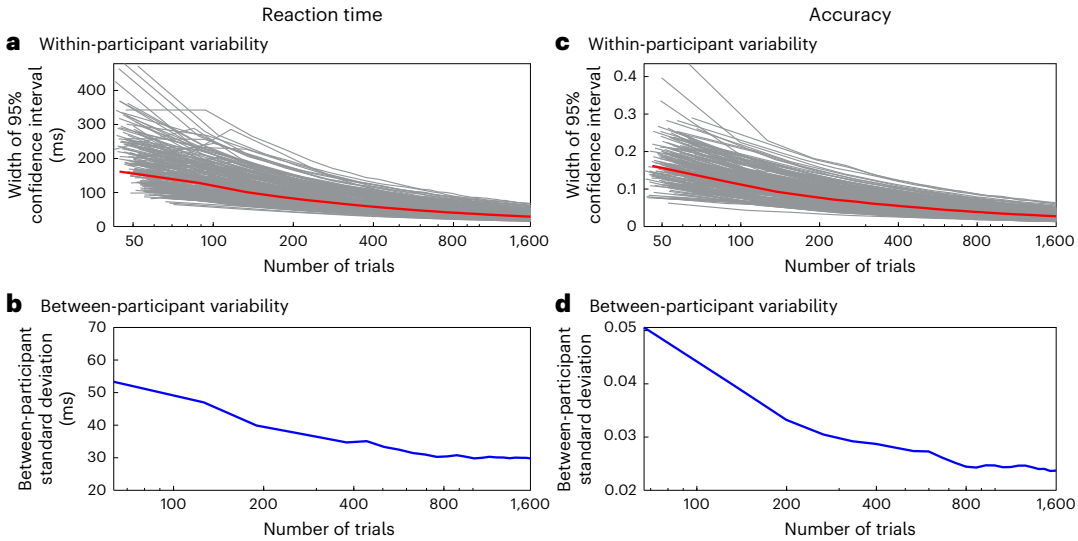


Fig. 4 | Repeated measures reduce both within- and between-participant variability. We used flanker task data from the work of Robinson and Steyvers[49], collected online through Lumosity. **a**, Stability curves for within-participant estimates of the congruency effect reaction time, based on the width of the 95% confidence interval of the mean congruency effect, estimated using bootstrapping (method 2). Method 1 is our preferred estimation method to plot the stability of congruency effects, as results are more interpretable in terms of the average performance outcomes one is likely to get across repeated data. However, it can be effectively implemented only with sufficiently large amounts of data to set the gold standard. Therefore, we used method 2 for this dataset. Each grey-shaded line is a single participant, and the overlaid red line is the median of the group. As in the EPIC dataset, within-participant error decreases with accumulating trials in the larger sample dataset of Robinson and Steyvers. Note that, overall, both within- and between-participant variability are higher in this dataset than in EPIC, perhaps due to online data collection, more heterogeneous group and/or wider age range; see the Supplementary Methods for details on the dataset. **b**, Between-participant standard deviation of the congruency effect reaction time plotted as a function of number of trials. Between-participant variability also decreases with trial numbers and reaches a stable point at ~1,000 trials. **c**, Stability curves for within-participant estimates of congruency effect accuracy. **d**, Between-participant standard deviation of congruency effect accuracy. The accuracy results are consistent with reaction time data in that both within- and between-participant variability decrease and stabilize at ~1,000 trials.

The simulation results in Fig. 5 show that Model 1 replicates key patterns observed in the empirical data of Robinson and Steyvers[49] (Fig. 4): When within-participant error is high with few trials, between-participant variability estimates are both biased and imprecise. Here, bias refers to a systematic deviation between the true and estimated parameter values, while imprecision reflects variability due to random error. Given that the stabilizing point of the between-participant standard deviation (~29 ms in Model 2) is our best true estimate, one can observe that the estimates under Model 1 are biased (inflated above 60 ms) and imprecise (as indicated by the large shaded error bars). With more trials, as the width of the 95% confidence interval for the congruency effect decreases within participants, the between-participant standard deviation decreases. By contrast, Model 2 features small within-participant error due to large trial sampling, and the between-participant standard deviation shows stable estimates. This suggests that, once within-participant error is reduced with large trial numbers, more accurate and stable between-participant variability can be revealed.

Besides collecting more trials, another effective way to correct for bias is to account for trial-level variability in the estimation of between-participant variability, as specified in Supplementary Equation 1 (refs. 33,39). This method (Extended Data Fig. 7) effectively corrects the inflation of between-participant variability, but imprecision of the corrected value can still be high (consistent with Rouder et al.'s[39] findings), probably due to imprecision in accurately estimating within-participant variance with small numbers of trials. Therefore, both unbiased and precise estimates of between-participant variation likely require at least an intermediate number of trials, even when corrected. However, this approach is helpful to apply in situations where group-level statistics are of interest but cannot provide precise estimates for an individual.

Within-participant error biases between-participant estimates

To further highlight the importance of minimizing within-participant error, we conducted additional simulations to examine how within-participant variability and sample size influence estimates of between-participant variability. This time we directly varied the size of within-participant variability, using parameters obtained from Hedge et al.'s[32] flanker task data. We set the true between-participant standard deviation to a fixed value and simulated participants with varying levels of within-participant standard deviation. We then measured the observed ('apparent') between-participant standard deviation for each within-participant standard deviation.

Figure 6a demonstrates that, although the true between-participant standard deviation is unchanged (18 ms), when within-participant standard deviation is higher than 9 ms (50% of the true between-participant standard deviation), the apparent between-participant standard deviation becomes both biased and imprecise. We replicated these results with different levels of between-participant standard deviation (Fig. 6b), and again, when the within-participant standard deviation is higher than about half of the true between-participant standard deviation, the apparent between-participant standard deviation starts to grow with higher error. Note that this growth is more prominent for smaller true between-participant standard deviations, indicating that measures with smaller true individual differences would be more severely affected by large within-participant errors. These results are consistent with Supplementary Equation 1: as the estimated between-participant variance is the sum of true between-participant variance and imprecision due to within-participant error, the results will be most affected by large imprecision when the true between-participant variance is relatively small.

In many cases, larger samples can compensate for the low reliability of individual data. However, increasing the number of participants

https://doi.org/10.1038/s41562-025-02198-2

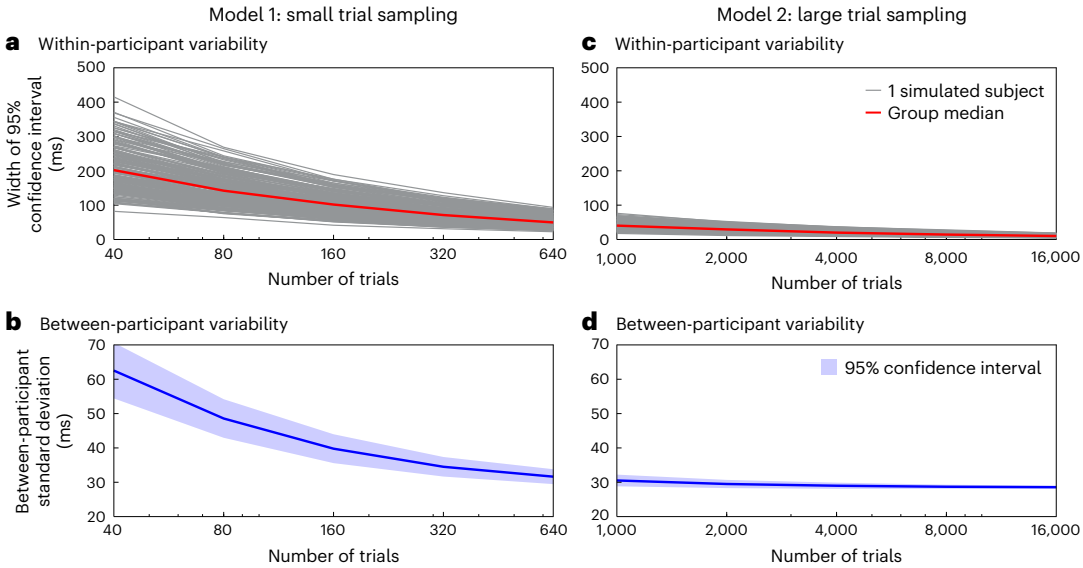


Fig. 5 | Between- and within-participant variability based on trial sampling. This figure shows the comparison of the two models that differ in the number of trials sampled from within-participant distributions to get the mean congruency effect: Model 1, with small trial sampling, sampled 40 trials per draw, while Model 2, with large trial sampling, sampled 1,000 trials per draw. These trial numbers were selected to match the NIH Toolbox Flanker Inhibitory Control and Attention Test (40 trials) and our estimate of the number of trials needed for stable congruency effects (1,000 trials). Furthermore, to examine how within- and between-participant variability change with increasing trials, the number of trials sampled was systematically increased from each starting point. We simulated 185 participants, whose parameters for within-participant distributions (mean, standard deviation, skewness and kurtosis) were based on the 185 participants from the work of Robinson and Steyvers[49] shown in Fig. 4 (see the Methods for more details on the simulations). **a**,**c**, The width of 95% confidence interval of the mean congruency effect across the number of trials (method 2). One grey line corresponds to one simulated participant, and the overlaid red line is the group median. **b**,**d**, The between-participant standard deviation of congruency effect plotted in a blue line and its 95% confidence interval as shaded error bars. Both within- and between-participant variability decrease with more trials in the small trial sampling and large error variance (Model 1) but are relatively unaffected by the number of trials in the large trial sampling and small error variance (Model 2), supporting that 1,000 trials can provide stable estimates.

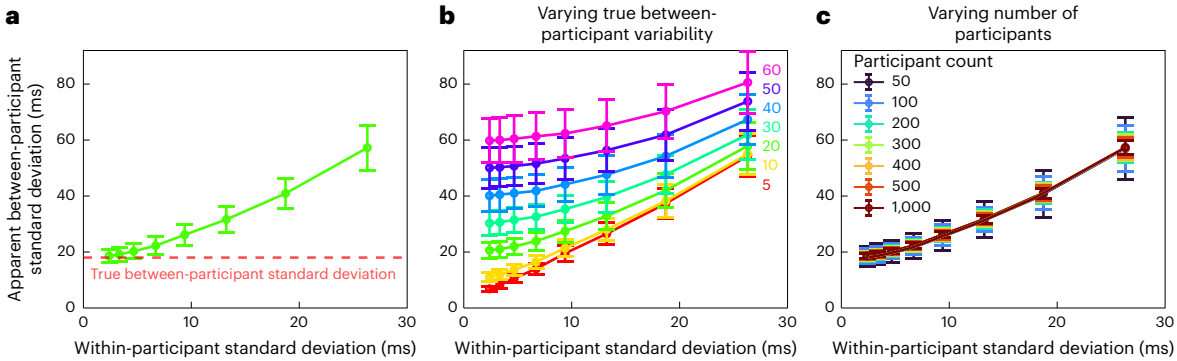


Fig. 6 | Within-participant variability inflates between-participant differences. a, A simulation of apparent between-participant standard deviation across different levels of within-participant standard deviation (2.4, 3.3, 4.7, 6.6, 9.3, 13.2, 18.8 and 26.3 ms; these values correspond to within-participant variability when trial number systematically varies from 6,400 to 50 in our dataset). The true between-participant standard deviation was set to a fixed value of 18 ms (marked by dashed red line) as in Hedge et al.'s[32] dataset. The simulation was repeated 1,000 times, and 100 participants were simulated in each simulation (see the Methods for details). The mean of 1,000 simulations is plotted with its 95% confidence interval as error bars. While the true between-participant standard deviation is unchanged, its apparent value increases with the within-participant standard deviation, indicating that large within-participant error inflates measures of between-participant differences. This contamination is most evident once the within-participant standard deviation is higher than ~9 ms, half of the true between-participant standard deviation. **b**, The simulation from **a** was repeated with varying levels of true between-participant standard deviation (5, 10, 20, 30, 40, 50 and 60 ms shown in different colour lines on the plot). Note that the smaller the true between-participant standard deviation, the more it is affected by increases in within-participant standard deviation. **c**, The simulation from **a** was then repeated with varying numbers of simulated participants (50, 100, 200, 300, 400, 500 and 1,000). Critically, increasing the sample size does not rectify large within-participant variance contaminating apparent between-participant standard deviation.

will not rectify bias in between-participant standard deviation. Figure 6c shows a simulation in which the number of simulated participants varies while the between-participant standard deviation is held constant. Although the error bars decrease with larger sample sizes, the overall pattern remains consistent, indicating that the inflation of between-participant variability cannot be resolved by increasing the number of participants. The within-participant variability itself needs to be reduced, such as through repeated measures. This fact is also apparent from Supplementary Equation 1, given that the number of participants does not affect between-participant variance.

This series of simulations demonstrates that, when within-participant variability is not appropriately addressed, measures of

https://doi.org/10.1038/s41562-025-02198-2

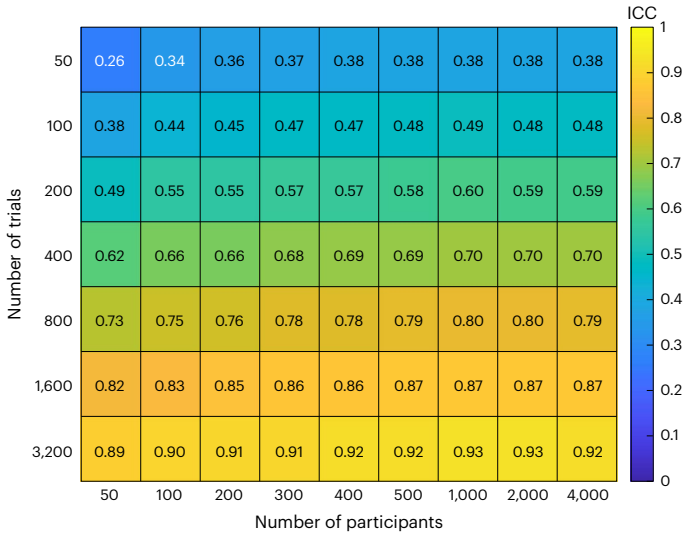


Fig. 7 | Effect of trial and participant numbers on ICC. This heatmap displays the ICC for the rank orders between the true mean and the apparent mean based on simulated data. A simulated participant's true mean congruency effect is the value directly sampled from a between-participant distribution, while the observed or apparent mean is the mean of a given number of trials sampled from a within-participant distribution (see the Methods for details). The two key manipulations were the number of participants drawn from the between-participant distribution (50, 100, 200, 300, 400, 500, 1,000, 2,000 and 4,000) and the number of trials drawn from the within-participant distribution (50, 100, 200, 400, 800, 1,600 and 3,200) to examine whether increasing either the number of participants or the number of trials improves the correlation between the true mean and the apparent mean. Increasing the number of trials alone results in excellent reliability above $r = 0.8$, whereas increasing the number of participants has relatively more limited effects.

between-participant variability will be incorrect[39,58]. Many past studies have measured inhibitory control using a small number of trials, likely resulting in within-participant variability that exceeds the expected between-participant variability. As a consequence, estimates of between-participant variability in the literature are likely to be often inflated and imprecise.

Trial counts matter more than participants in reliability

Given that increasing the number of participants can have limited benefits for accurately estimating between-participant variability, we compared the effects of increasing the number of participants versus the number of trials on cross-participant reliability, measured with intraclass correlation (ICC). This is an important consideration when determining the sample size of a study with limited resources, especially given recent examinations of the trade-off between participant numbers and testing duration in correlational studies[56]. Here, we calculated the correlation coefficient of the rank orders between a simulated participant's true mean congruency effect and the apparent mean across varying numbers of participants and trials using Hedge et al.'s[32] flanker task data (Fig. 7). Replicating and extending the findings from Fig. 6c, we show that collecting more trials—rather than increasing the number of participants—leads to excellent cross-participant reliability.

In this simulation, increasing the number of participants is most effective in improving reliability when the sample size is small (<200 participants). The effect becomes negligible when the number of trials is beyond 800. For example, the ICC is 0.87 when the number of trials is 1,600 and the number of participants is 500, and it remains unchanged even when participant numbers increase to 4,000. Even with thousands of participants, the ICC can be below 0.5 when individual-level estimates of the congruency effect are imprecise due to having trials as small as 50.

In contrast, increasing trials effectively improves the ICC. By increasing the number of trials, it is possible to achieve excellent reliability in even relatively small samples—indeed, with sufficient trials (>1,000), even with 50 participants, the reliability is above 0.8. For example, with 3,200 trials and 50 participants, the ICC is 0.89. This contrasts with the case of having 4,000 participants with only 50 trials each, where the ICC is 0.38.

Consistent with Fig. 6c, these simulation results demonstrate the importance of acquiring precise individual estimates by sufficient trial sampling rather than participant sampling. These results convey a critical message regarding the choice between expanding the number of participants versus the number of trials when resources are limited, especially when examining measures with high within-participant variability such as inhibitory control. We expand on this further in the Discussion.

Advanced models benefit from more within-participant data

We have demonstrated the necessity of repeated measures to obtain precise individual estimates and accurately assess individual differences in classic congruency effect measures. In our final analyses, we explored the implications of sampling on alternative methods to analyse inhibitory control in congruency effect paradigms. Advanced modelling approaches, such as drift–diffusion modelling[36,46,60], factor analysis[47] and Bayesian hierarchical modelling[33], have been proposed to improve reliability or investigate unobserved variables in task performance. These methods are often implemented with fewer than a hundred trials[36,61]. Building on our findings regarding the critical role of sufficient sampling, we evaluated how varying trial numbers affect the robustness of these advanced modelling approaches.

We first show results of EZ-diffusion modelling[62] using flanker task data of Robinson and Steyvers[49]. We calculated the split-half reliability of the modelling parameters across different numbers of trials.

Figure 8 demonstrates that increasing the number of trials improves the ICC for both the congruency effect and the modelling parameters. Specifically, the ICC for the modelling parameters mirrors changes observed in the ICC for reaction time and accuracy of the congruency effect. The ICC for the drift rate increases systematically with a higher number of trials, requiring at least 800 trials to reach an ICC of 0.8. In most cases, the ICC for the drift rate does not exceed that of reaction time or accuracy. Therefore, the modelling results are expected to achieve high reliability once congruency effects are measured with high precision, such as with sufficient sampling. We also calculated the bootstrapped 95% confidence interval for the ICC of the modelling parameters. Extended Data Fig. 8 shows that, as the ICC improves with more trials, its precision also increases, as indicated by the narrowing of the error bars. We replicated this pattern using Hedge et al.'s[32] data (Extended Data Fig. 8) and a simulation of a larger number of trials (up to 3,200 trials; Supplementary Fig. 12).

In our confirmatory factor analysis (CFA; Extended Data Fig. 9), we similarly observed that the reliability of the modelling outcomes is constrained by the reliability of the original data. We simulated three congruency tasks data using parameters from Hedge et al.'s[32] flanker and Stroop tasks, as well as the EPIC prime–probe task, and ran CFA assuming a single shared latent factor. We examined the reliability of the factor scores across varying trial numbers. The results show that the reliability improves with more trials, and notably, cross-task correlation plays an important role: if cross-task correlation is weak, even with sufficient sampling, the reliability of the factor score can remain low. Conversely, with high cross-task correlation, the reliability of the factor score exceeds that of the individual task congruency effects. Thus, precise individual measures can influence CFA results, particularly in cases where cross-task correlation is modest. This finding is not entirely surprising, given that our CFA model assumed a single factor (based on prior P-technique factor analysis in the EPIC dataset, which suggested that all participants had at least one estimable latent factor

https://doi.org/10.1038/s41562-025-02198-2

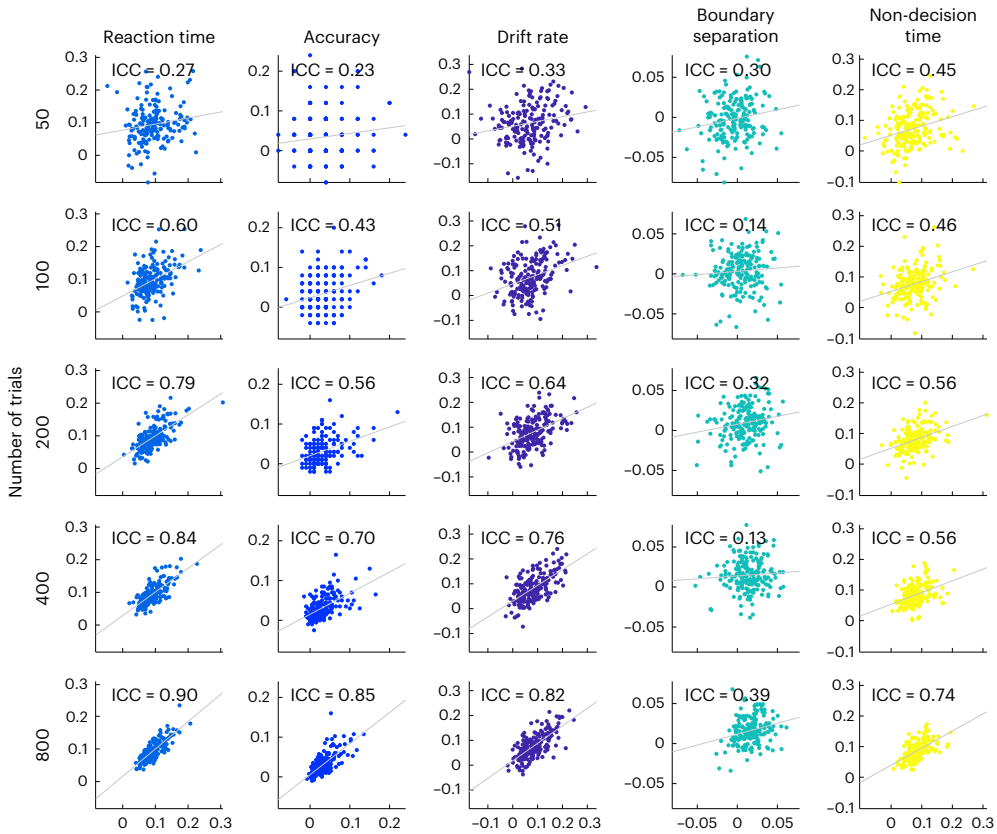


Fig. 8 | Reliability of EZ-diffusion modelling across increasing trial numbers. We used 185 participants from the dataset of Robinson and Steyvers[49]. As EZ-diffusion modelling requires inputting both reaction time and accuracy[62], we compared the split-half reliability of the modelling parameters (drift rate, boundary separation and non-decision time) to the congruency effect reaction time and accuracy. All results are difference scores between congruent and incongruent trials; for accuracy, drift rate and boundary separation, incongruent trials were subtracted from congruent trials. For the rest, congruent trials were subtracted from incongruent trials. We systematically increased the number of trials (50, 100, 200, 400 and 800) and calculated the cross-participant reliability with ICC. The results show that the reliability of the drift rate improves with more trials, but it seems to be limited by the extent to which the reliability improves for reaction time or accuracy. See Extended Data Fig. 8 for the same plot but using Hedge et al.'s[32] data and Supplementary Fig. 12 for simulation results with extended sampling.

that the three tasks load on). However, this dependence emphasizes the importance of selecting appropriate priors for factor-level structure to ensure reliable factor analysis results.

Finally, we compared frequentist non-hierarchical and Bayesian hierarchical methods of analysing the data from the work of Robinson and Steyvers[49] (Extended Data Fig. 10). Bayesian hierarchical models correct for measurement imprecision in conditions of low trial-numbers by shrinking unstable parameter estimates towards the group mean. This has the effect of downwardly correcting the error of between-participant variance estimates[33]. Our own simulations demonstrate that a Bayesian hierarchical approach provides large benefits only when the number of trials is small and within-participant variance is large. In other words, if one has to accept some degree of imprecision due to resource constraints and a small dataset, then a Bayesian hierarchical approach will provide less imprecision in the parameter estimates than a non-hierarchical approach. However, both approaches require a similar number of trials to obtain parameter estimates precise enough to produce mean absolute differences in the range of 4–9 ms from the true values. Thus, while alternative analysis strategies will probably be useful for improving inference in cases with high within-participant noise, these strategies are themselves improved by having more data. Benchmark datasets, such as EPIC, can help to establish better priors for improving implementations of these analysis and guidelines for the number of trials necessary to obtain different levels of precision.

Discussion

We have empirically quantified the number of trials necessary to obtain precise estimates of the congruency effect. We collected a dataset with more than 5,000 trials for each of our 9 participants in 4 different tasks, 3 of which probed the congruency effect. Using this dataset, we demonstrated that within-participant variability of the congruency effect can be significantly reduced through extended sampling, with gains plateauing around 1,000 trials (500 trials per experimental condition). Our work complements prior findings from simulations based on more limited empirical examinations[39,63,64] and provides concrete estimates of the trial numbers needed for different levels of reliability.

We replicated and expanded our findings with two additional public datasets[32,49], showing that within-participant variability is of central import in the comparison of inhibitory control across people, as high levels of within-participant error systematically contaminate estimates of between-participant differences. This error cannot be resolved by increasing participant numbers and persists in alternative analysis methods, including drift–diffusion modelling, factor analysis and Bayesian hierarchical modelling.

Jointly, these results suggest that additional attention to within-participant variability is warranted in the study of inhibitory control for both basic research and clinical applications. Although here we focused on the congruency effect, our findings on the contamination between within- and between-participant variability and how the

number of trials reduces bias and imprecision in between-participant variance are likely to apply to other tasks and measures.

We share our EPIC dataset as a benchmark resource for the investigation of inhibitory control. The dataset can serve as a tool to understand sources of variability in inhibitory control and to quantify changes in task performance over time. Our dataset can also provide means to examine relationships across different task measures and benchmark methods for analysing congruency effect tasks. These ideas are explored in more detail below.

Inhibitory control estimates can be precise

As laboratory paradigms to investigate inhibitory control have been scrutinized for having low reliability[32], several alternative approaches have been suggested. Some have proposed moving away from difference score metrics that contrast trial conditions[31,65–68]. However, this modification can alter the interpretation of the congruency effect, as non-subtracted metrics can be influenced by other aspects of processing beyond the intended comparison (for example, see Extended Data Fig. 2). Others have proposed improving task designs to increase control demands or arousal, for example, by combining different congruency tasks or using gamification[45]. Utilizing smartphones to access a broader population[44] or selecting high conflict trials[69] have also been suggested. The use of hierarchical models has been particularly recommended for their ability to separately model trial noise to better estimate true between-participant variability[33,39,48,63,64,70].

Alongside these approaches, our results suggest that large trial numbers are beneficial for more precisely estimating both within- and between-participant variability. Several studies have previously noted that collecting more trials would give better estimates of the congruency effects[39,44,64]. For example, Rouder and colleagues[39] reviewed published studies on the typical ratios of between- to within-participant variability of the congruency effect and noted that, given the small ratios, most studies collect too few trials. While there are concerns that more testing could introduce systematic variability due to fatigue or boredom[30], our study demonstrates with empirical evidence that approximately 500 trials per condition yield precise individual-level estimates, with relatively small influences from performance impairment over time.

Indeed, using our EPIC dataset, we demonstrate the efficacy of precision approaches in improving reliability even in alternative modelling approaches. These approaches are particularly effective when the proportion of error variance to total variance is large (or signal-to-noise ratio is low[71,72]). These measures from our extensive dataset can serve as gold-standard empirical estimates of congruency effects within individuals to benchmark the effectiveness of various methods to improve reliability in future studies.

EPIC, a dataset for inhibitory control research

Repeated testing enables the measurement of other sources of within-participant variability, such as practice effects. All participants showed congruency effects throughout the duration of data collection, but decreases in the magnitude were observed with additional experience performing the tasks in reaction time data (for prior studies also reporting similar effects, see ref. 73–75). These performance improvement effects were absent in the data from the work of Robinson and Steyvers[49]. Whereas their data were collected over the span of years, our participants performed each task twice a week for 9 weeks. Accordingly, we speculate that the length of interval between sessions may play a critical role in observing these effects.

One future avenue of interest is to investigate the properties of these temporal effects and how they vary across tasks, measures and individuals. Performance improvement effects were most prominent in the flanker task and least so in the Stroop task, which had the smallest and largest within-participant variability, respectively, among the three tasks (Supplementary Fig. 5). Participants with the smallest within-participant variability also demonstrated the most prominent improvement effects across all three tasks (for example, EPIC 03 and 07). Note that EPIC 03 also showed overall performance improvements, as reaction time was shorter and accuracy was higher in later sessions (Extended Data Fig. 4). Accordingly, it may be that participants with better task engagement, reflected in smaller measurement error, are more likely to acquire these effects. This hypothesis warrants further investigation in future studies.

Examining other sources of within-participant changes across and within sessions related to repeated testing revealed no critical signs of performance deterioration in our dataset. However, we observed some variability across sessions, with the extent of variability differing by task and participant. Small effects of fatigue were also observed in the Stroop task over the 1 h of testing. Hence, using our dataset to explore day-to-day and within-session variation (for example, through sliding window analyses) could offer valuable insights into the nature and sources of variability in inhibitory control. For example, if the congruency effect remains relatively stable within an individual (not exceeding variation expected from sampling variability), this may suggest that it has relatively trait-like characteristics. Variation over time within a session or across sessions, conversely, could be linked to differences in states, arousal, or other individual characteristics. Longer-term dynamics may also be of interest for tracking longitudinal changes in inhibitory control, particularly in relation to developmental and degenerative conditions. Our dataset should serve as a valuable step to initiate this exploration.

Precise estimates are needed for individual differences

The precision of individual-level data should not be compromised even in research focused on individual differences. We demonstrated through simulations that within-participant and between-participant variability are linked. This leads large within-participant variability to systematically contaminate estimates of individual differences. Our simulation results are consistent with past findings[33,39] and converge with the expectations from statistical equations of between-participant variance (Supplementary Equation 1).

Although we examined these effects in the domain of behavioural measures of inhibitory control, they are likely to extend broadly to other behavioural and neural measures. Indeed, the properties of our relatively simple simulation (Fig. 6) and mathematical expectations from statistical inference (Supplementary Equation 1) suggest that estimates of between-participant variability will be contaminated in any situation where within-participant variability is high and cannot be recovered by simply sampling more participants. Thus, these results call for enhanced attention to the precision of individual-level estimates across a range of experimental paradigms, particularly in large-scale studies designed to understand individual variation.

Clinical applications, in particular, are likely to be impacted by these findings. Deficits in inhibitory control have been linked to ageing[76], psychiatric disorders[77,78] and maladaptive behaviours in daily life, such as suicide[79] and drug addiction[80]. However, research in these areas often yields inconsistent findings. Some studies have found non-significant differences in inhibitory control between psychiatric groups and healthy controls[81] or poor correlations between laboratory tasks and self-reported surveys that are defined to measure similar constructs of inhibitory control[82]. As a result, one might conclude from these studies that older adults or psychiatric patients do not have deficits in inhibitory control[83] or that the laboratory tasks are not valid measures of real-world inhibitory control[81,82].

However, instead, it could be that inhibitory control measures were imprecise at the level of individual participants. Study designs for examining within-participant effects of experimental manipulations have come to prominence because of their ability to measure robust group-level effects, including the congruency effect[21,31]. This effect was designed to be robust on average across participants but not at the level of individuals. Indeed, many past studies have acquired about

a hundred trials or less per participant in an effort to collect a large group sample (for example, refs. 36,76). The large amount of error in individual estimates of inhibitory control with these trial numbers probably hampers our ability to interpret these measures and connect them with other constructs, such as measures of psychopathology.

Having precise behavioural measures is also critical for linking them with neural characteristics. Substantial effort and resources have been dedicated to understanding how interindividual variations in behavioural phenotypes are associated with individual differences in neural features through large consortium datasets[17,84,85]. However, studies based on the ABCD and HCP datasets suggest that brain–behaviour links are weak[86,87]. Supporting this, Kong and colleagues[88] demonstrated that brain-based prediction accuracy for flanker task performance was particularly low among the 58 behavioural measures in the HCP dataset. Gell and colleagues'[41] simulation study helps to explain this, showing that predicting behavioural measures with low reliability from brain features results in extremely low prediction accuracy. Given our findings, poor predictive accuracy is not surprising for datasets and consortia based on direct or modified versions of the NIH Toolbox, which typically includes only 40 trials per participant. Therefore, highly reliable individual measures, such as through sufficient trial sampling, are likely to improve predictions in future efforts to connect brain and behavioural measures of inhibitory control[89,90].

Precise estimates advance executive function research

Executive functions are a set of cognitive processes that regulate thoughts and behaviours to achieve goals, with inhibition being one of the key components. Extensive prior research has defined, characterized and classified executive functions to demonstrate their central role to function adaptively in daily life[3,38,91]. However, a crucial question remaining is the unity and diversity of executive functions: whether control is a unified concept, and whether different tasks measure similar or related control functions[38]. Some have argued that inhibition is not a coherent theoretical entity and that different inhibitory control tasks may tap on different constructs[33]. Others have argued that inhibitory control reflects a unitary phenomenon with evidence that a shared latent inhibitory control factor exists across tasks[38,92]. Factor analysis is a tool for removing error variance and examining underlying shared constructs[93]. Our simulation results showed that factor analysis is also affected by trial variability, highlighting the need for sufficient testing to ensure reliable modelling. Thus, obtaining robust individual-level measures through sufficient sampling is important for future studies aimed at improving understanding of executive functions and their relationships to various measures.

Collecting more trials versus more participants

Despite prior attention to reliability[30,32,41,58], collecting small numbers of trials per participant remains common practice, as promoted by tools such as the NIH Toolbox and their use in large consortium datasets such as the ABCD and HCP. In applications related to brain–behaviour associations, substantial attention has been paid to issues of participant number in improving reliability[87], but relatively less to issues of within-participant data collection[89,90].

Our simulation results indicate that simply having more participants when their individual-level estimates are biased will continue contaminating estimates of true variation across individuals. Large consortium studies often collect fewer than a hundred trials from hundreds or thousands of participants (for example, ref. 94). Our results suggest that this approach probably leads to erroneous estimates of individual differences, especially for measures that require detection of small differences. Rather, having sufficient trials (~1,000 trials) provides precise individual estimates and reveals true individual differences even without necessitating hundreds of participants. Of course, large participant samples are still necessary to represent the population and to generate inferences about individual difference variables[95,96]. However, our findings imply that collecting data from large samples at the expense of individual precision may be a misleading strategy, leading to systematically incorrect estimates of individual variability.

Deciding on the optimal combinations of trial size and number of participants is a challenging enterprise when resources are limited. However, our results suggest that increasing participant numbers at the expense of very low trial numbers per participant can have concerning consequences on measurement. To be concrete, the simulation in Fig. 7 shows that having 50 participants with 800 trials achieves a similar ICC (0.73) to having 1,000 (or 4,000) participants with 400 trials. Thus, assuming 1,000 trials are collected per hour, one can achieve similar reliability with 40 h of testing compared with 400 h. More dramatically, one can achieve two times the ICC by going from 4,000 participants with 40 trials (200 h of testing) to 50 participants with 800 trials (40 h of testing).

Guidelines for precise estimates of inhibitory control

We have shown that the stable congruency effect estimates emerge with more than 1,000 trials. One might ask whether it is feasible to collect more than 1,000 trials per participant. It took less than an hour to collect 1,000 trials for our study.

Additional time per participant can be a constraint for studies that require a large sample of participants or multiple tasks. Moreover, for neuronal measures, such as functional magnetic resonance imaging, additional design constraints may exist that can lengthen the duration of each trial, such as allowing for sufficient intertrial intervals to return to baseline. Another critical problem is recruitment: it may be easy to recruit a large number of participants for studies taking only a few minutes, while it could be more difficult to recruit them for studies taking hours.

Despite these constraints, our findings demonstrate the importance of precise individual data. Imprecise individual estimates do not guarantee reliable examination of individual differences even with collecting hundreds of participants. In this vein, several other approaches have also been suggested to improve congruency effect measurement, including drift–diffusion modelling[46], factor analysis[47] and Bayesian hierarchical modelling[33]. We examined the role of trial numbers in each of these strategies and demonstrated that, in all cases, precision is improved with higher amounts of data per participant.

Therefore, while the precision approach of collecting large per-participant data represents more time on tasks than is currently carried out in many studies, we believe the added precision is worth the investment. This may lead researchers towards study designs with fewer tasks per participant in an effort to obtain more reliable estimates in the individual measures. Alternatively, researchers may seek to wed smaller, extensive sampling datasets with larger-scale studies[97,98], for example, by using small datasets to improve priors in the analysis of noisier larger-scale datasets[90].

Conclusion

Using a dataset with extensive per-participant data, we have demonstrated that it is possible to obtain highly precise congruency effects. This dataset provides a valuable resource for testing and validating methods to examine inhibitory control. Our findings, supported by both empirical and simulation data, consistently highlight the importance of extended sampling in obtaining precise individual-level estimates and reliable between-participant differences. These principles may extend beyond inhibitory control, with broad implications for improving the reliability of measures in both clinical and cognitive neuroscience research.

Methods

Overview and datasets

Our goal was to determine whether we could obtain highly precise individual estimates of inhibitory control with sufficient sampling.

Thus, we collected new data, which we term the EPIC dataset. This dataset includes extended amounts of data from 9 individuals who were tested on 4 inhibitory control tasks across 36 days (2 tasks per day; 18 days per task). Other daily measures were also collected, including sleepiness (Stanford Sleepiness Scale), sleep time, anxiety (State-Trait Anxiety Inventory), mood (Positive and Negative Affect Schedule), depression (Beck Depression Inventory), substance use and simple reaction time. We also measured the Big Five personality traits. We describe the dataset and measures in more detail below. We have made this dataset a public resource for the community to test a range of questions regarding inhibitory control, within-participant variation, practice and relationships across measures. For the purposes of this study, we focused on the precision of the congruency effect measures of inhibitory control, based on three of the four inhibitory control tasks.

Our preliminary analyses focused on establishing the reliability of the congruency effect in the EPIC dataset under different conditions. We then replicated findings from two public datasets of 495 participants in a flanker task[49] and of 112 participants in a flanker and Stroop task[32]. These datasets are from a larger sample of participants, although with more limited amounts of data per participant. These secondary datasets are described in more detail in the Supplementary Methods. We used these datasets to establish properties of between-participant variability relative to within-participant variability. We complemented our work on these three datasets with simulations, as described in more detail below.

EPIC dataset

Description. The EPIC dataset includes data from three congruency tasks: a flanker task, a prime–probe task and a Stroop task. These tasks measure the congruency effect in diverse ways, tapping differences in stimulus presentation, spatial and temporal adjacency of distractors to targets and loci of interference or control[92,99,100]. The flanker task instigates interference between a target and distractors in space, whereas the prime–probe task induces conflict in time between a prime (distractor) and a probe (target), which is why some consider this task to be a temporal flanker[101]). In the Stroop task, conflict arises when the target and the distractor are associated with different attributes of the same stimulus, requiring suppression of a prepotent response to the distractor. We also tested a go/no-go task, which measures rapid inhibition of motor execution, but did not include them in our analyses of investigating precise estimates of congruency effects (see Fig. 1, a visual abstract, for the description of the dataset). All data are available at https://osf.io/jk9nb/, and the experiment and analysis code are available via GitHub at https://github.com/GrattonLab/LeeSmith_EPIC.

Participants. Data were collected from nine healthy adults who were either members of the laboratory or students at Northwestern University (age mean 25 years, standard deviation 3.61 years; 5 females and 4 males). The study was approved by the institutional review board of Northwestern University (STU00211073). All provided written informed consent to participate. Participants were compensated for each session and received a completion bonus upon completing all 36 sessions. All had normal or corrected-to-normal vision. Participants were either tested in the laboratory or at home by taking the laboratory computer home. Data from one participant were removed due to an issue related to key release resulting in consecutive error trials in later sessions of the prime–probe and Stroop tasks (EPIC 09). The Article results are based on the remaining eight participants.

Apparatus. Participants were seated approximately 60 cm away from an LCD monitor. The screen was set to have a resolution of 1,440 × 900 pixels and a refresh rate of 60 Hz. All experiments were programmed with MATLAB (www.mathworks.com) and Psychtoolbox (3.0.16). Responses were collected with a standard computer keyboard.

Data acquisition procedure. Participants each completed 36 sessions with 4 sessions each week. In each session, participants performed two of the four inhibitory control tasks (flanker, prime–probe, Stroop and go/no-go tasks), and the order was pseudorandomized and counterbalanced across participants. The order was miscollected for EPIC 10, so the data collection was incomplete for some tasks upon completing the 36th session; this participant completed two additional sessions to reach the same level of task completion. In the beginning of each session, participants completed a 5-min survey on Qualtrics, responding to questions about their mood, current emotions and activities from the previous 24 h. The survey was followed by a simple reaction time task consisting of 25 trials per day. In this task, participants were asked to press the space bar upon seeing a white square appearing in the centre of the screen. The purpose was to measure reaction time for motor execution in response to a stimulus presentation. The data were not analysed but are released for interested users.

Task design and procedure

Flanker task. In each session, participants had 24 trials of practice, followed by 4 blocks with 100 trials per block of the main experiment. The task was to respond to the direction in which a central arrow points, while ignoring the arrows presented beside it, by pressing the left or right arrow key. Participants were instructed to respond as quickly and accurately as possible. Participants received feedback on every trial during the practice and only at the end of each block for the main experiment. Each trial started with the presentation of a white central fixation cross on a black background for 1,500 ms. This cross was followed by a display of one target arrow in the centre and four flanker arrows, two placed on the left and the other two placed on the right of the target, all in white colour for 500 ms. The keyboard response was recorded from the onset of the stimulus until the end of the following fixation cross period. A trial was considered congruent when the target arrow and the flanker arrows pointed in the same direction and incongruent when they pointed in opposite directions. The number of trials was equal between the two conditions. The final total number of trials collected per participants is 7,200.

Prime–probe task. The task design was based on Weissman et al.'s[102] study. Participants completed 24 trials of practice followed by 4 blocks of 96 trials in each session. The task was to ignore the preceding prime letter and respond to the probe letter. One of four letters was presented to which participants responded with their right hands (by pressing the key '1' in response to letter 'A', '2' to letter 'B', '3' to letter 'Y', and '4' to letter 'Z'). Participants were instructed to respond as quickly and accurately as possible. Participants received feedback on every trial during practice but only at the end of each block during the main experiment. A white fixation cross, presented in the centre of the black screen for 1,067 ms, was followed by a white prime letter, which was presented for 200 ms. After a blank screen appeared for 33 ms, the probe letter, also in white, was presented for 200 ms. The keyboard response was recorded from the onset of the prime until the end of the following fixation cross period. Trials were considered congruent when the prime and the probe letters cued the same response and incongruent when they did not. Each trial type (letter combination) was presented an equal number of times; thus, there were an equal number of congruent and incongruent trials. The final total number of trials collected per participants is 6,912.

Although not analysed in this Article, the prime–probe task was designed to examine how the congruency effect on the current trial is influenced by the congruency status of the previous trial[103], without feature integration and contingency learning[53,104]. To avoid contingency bias, the four letters were grouped into two sets (A and B; Y and Z) and stimulus response repetitions were prevented by switching between sets on each trial. The trial sequence was generated under the

constraint that each congruency sequence (cc, ci, ic and ii) occurred an equal number of times. However, because the first trial of each block does not have a preceding trial, one of the four sequences in each block occurred one trial less than the others. Across blocks, the sequences were balanced, with each of the four sequences serving as the less frequent sequence in one block.

Stroop task. Participants completed 25 practice trials, followed by 4 blocks of 108 trials in each session. The task was a colour–word Stroop task, where participants responded to the colour of the word rather than its meaning. The colours were red (correspondingly pressing the key, '1'), yellow ('2'), green ('3') or blue ('4'). Participants responded with their right hands. Participants were instructed to respond as quickly and accurately as possible. Participants received feedback at every trial during practice and only at the end of each block during the main experiment. Every trial started with a white fixation cross presented for 1,000 ms on a black screen. A word followed this cross and was presented on the screen for 1,000 ms. On congruent trials, the ink colour and word meaning were the same, whereas on incongruent trials, they were different. One-third of the trials were neutral conditions during which words irrelevant to colour ('dog', 'bird', 'horse' and 'cat) were presented. With the neutral condition, one can examine the facilitation effect (neutral-congruent) and the interference effect (incongruent-neutral). We chose to have the neutral condition only in the Stroop task because our primary focus was to maximize the likelihood of obtaining precise estimates in the common inhibitory control paradigms, specifically the congruency effect. The downside of including a neutral condition was having fewer trials per condition. The numbers of trials across the three experimental conditions were the same. The final number of trials collected per participant is 7,776 (5,184 trials for calculating the congruency effect, excluding the neutral trials).

Go/no-go task. The task design is a modification of Redick et al.'s[105] paradigm. In each session, participants completed 20 trials of practice. The main experiment was composed of 4 blocks of 100 trials. The task was to respond by pressing the 'x' key with their right index finger when the letter 'X' appeared (go trials) and withhold responses when other letters ('B', 'C', 'F', 'G', 'H', 'J', 'K', 'P', 'T' or 'Z'; no-go trials) appeared. Participants were instructed to work as quickly and accurately as possible. Each trial started with a white fixation cross appearing at the centre of the black screen for 700 ms. This cross was followed by a white letter presented for 300 ms. Only 20% of the trials were no-go trials to induce prepotent response execution in the frequent go trials. The total number of trials collected per participant is 7,200 trials. This task was not analysed here, but the data are made available.

Data analyses

Preparing EPIC data for analyses. To calculate the mean congruency effect of 18 sessions and the grand mean of all sessions, we excluded outlier trials that deviated by more than three standard deviations from the mean within each experimental condition. We used the violinplot.m function in MATLAB to draw violin plots in Fig. 2 and Supplementary Figs. 1–4 (ref. 106). Before examining the stability of congruency effect results (for example, Fig. 3), we dropped the initial two blocks (200 trials for flanker, 192 trials for prime–probe and 216 trials for Stroop) of the first session to reduce errors associated with learning the task rules and stimulus–response mappings. We also regressed out the improvement effects in reaction time within each task (Extended Data Fig. 3); after plotting the mean congruency effect as a function of growing number of trials (by progressively adding half a block of trials), we fit a simple linear model (see Extended Data Fig. 6 for before-and-after linear regression). The residuals of this model were used for analyses. For completeness, we also plotted stability curves without regressing the improvement effects (Supplementary Fig. 5).

Within-participant precision of the congruency effect. We used two methods to measure the within-participant precision of congruency effect estimates. Method 1 served as the primary approach for datasets with a sufficient number of trials per participant, as it assesses replicability in independent samples within a participant. For each participant, data were divided into small units (that is, half a block, ~50 consecutive trials per unit). Randomly, half of these units were assigned to a reference set, totalling 2,592–3,600 trials. This reference set provided the best estimate of the true congruency effect score for each participant. To determine the sampling size that gives a comparable estimate to this reference score, one unit from the rest half was randomly selected with replacement and progressively added to a test set sample. The absolute difference between the test set sample's congruency effect and the reference set's congruency effect was then calculated. This procedure was repeated until the test sample size was comparable to that of the reference set. We repeated this process 5,000 times with different splits of the participant's data into reference and test samples. The final results plot the mean across these 5,000 repetitions.

Method 2 was the approach used in datasets with insufficient number of trials for test–retest comparisons and served to replicate the findings of method 1. In this method, each participant's data were also divided into small units (~50 trials), which were randomly selected and added to a growing sample. At each step of adding a unit, the mean congruency effect of the growing sample was calculated. This procedure was repeated 5,000 times, producing 5,000 estimates of the mean. These data were then used to calculate the 95% confidence interval of the mean. The width of the confidence interval was used as an estimate of within-participant variability. The correlation coefficients between the two methods of estimating precision in the congruency effect for each participant for each task were $r \geq 0.98$. Note that these same approaches were used for examining the within-participant reliability of reaction time (Fig. 3 and Supplementary Fig. 5), accuracy (Supplementary Fig. 7) and IES (Supplementary Fig. 8). A subset of these approaches was used to replicate the findings in our secondary datasets (Fig. 4).

In addition to the small segments used for the figures in this Article, we tried splitting data into bigger segments (for example, ~400 trials instead of 50 per unit). Trials are collected consecutively in experiments, and so shared error variance across trials may exist. Splitting data into units that are too small and randomly sampling them may give estimates of necessary numbers for stable results that are overly optimistic. Results showed that, while the within-participant variance slightly decreased with smaller segments for some participants, the differences in trajectories were trivial, suggesting similar stabilization points across different segment sizes (Supplementary Fig. 11).

Simulations

We ran simulations to further investigate the associations between within- and between-participant variability and the ability of different analysis methods to address these associations.

Simulation 1—effects of trial sample size on estimate variability. To investigate the hypothesis that between-participant standard deviation stabilizes with sufficient trial sampling when inflated by large within-participant error, we conducted and compared simulations of two models (Fig. 5). For both models, data were simulated on the basis of the selected sample of 185 participants from the dataset of Robinson and Steyvers[49], each with more than 2,500 correct trials. Each participant's distribution was simulated using the Pearson system, which constructs a distribution based on input parameters for mean, standard deviation, skewness and kurtosis[107]. Critically, the number of trials sampled per simulated participant differed between the two models; for the small trial sampling model (large within-participant variance), 40 random trials were sampled from the distribution, whereas for the large trial sampling model (small within-participant variance), 1,000 trials

were sampled. Forty were tested in the NIH Toolbox Flanker Inhibitory Control and Attention Test, while 1,000 trials is the amount we suggest provides stable estimates of the congruency effect. Furthermore, to observe how between-participant standard deviation changes with increasing trials, we systematically increased the number of trials drawn from each starting point (40, 80, 160, 320 and 640 versus 1,000, 2,000, 4,000, 8,000 and 16,000). The two models were created with 100 repetitions to get the mean and its 95% confidence interval across iterations.

Simulation 2—impact of within-participant variability on apparent between-participant differences. To examine when within-participant variability starts to contaminate apparent between-participant variability, a series of simulations was conducted (Fig. 6). First, based on Hedge et al.'s[32] study of 101 participants, we created a congruency effect distribution with a mean of 40 ms, standard deviation of 18 ms, skewness of 0.39 and kurtosis of 2.95. From this distribution, 100 simulated participants were sampled. For each simulated participant, we set the mean to the sampled value from the full distribution and the within-participant standard deviation to one of the preset values: 4.7 ms, 6.7, 9.3, 13.1, 18.8, 26.8, 38.1 and 53.2 (these values represent the 95% confidence interval of the mean congruency effect in the EPIC flanker task data, corresponding to 6,400, 3,200, 1,600, 800, 400, 200, 100 and 50 trials, respectively). The skewness of this within-participant distribution was −0.0022, and the kurtosis was 2.9967. We randomly sampled data from this simulated participant. The data were accumulated across 100 simulated participants, and the apparent between-participant standard deviation was measured. This simulation was repeated 1,000 times to plot the mean and its 95% confidence interval. We also conducted similar simulations but with varying between-participant standard deviation (5, 10, 20, 30, 40, 50 and 60 ms) and numbers of simulated participants (50, 100, 200, 300, 400, 500 and 1,000) to examine how the size of individual differences and number of participants affect the contamination.

Simulation 3a—effects of number of participants and number of trials on rank order. We used the dataset from Hedge et al.[32] to simulate cross-participant rank order consistency across different numbers of participants and trials per participant (Fig. 7). To set the parameters for these simulations, we first calculated the mean reaction time and accuracy for congruent and incongruent trials of 101 participants of Hedge et al.'s flanker task data (group mean 419 ms (congruent), 460 ms (incongruent); standard deviation 44 ms (congruent), 52 ms (incongruent)). As congruent and incongruent trials are highly correlated[30,31], we generated correlated samples using a multivariate probability distribution (MATLAB function, copulas). Each simulated participant's mean was drawn from this distribution. Next, each participant's distribution was generated using the within-participant standard deviations calculated from Hedge et al.'s data (group mean standard deviation 77 ms (congruent), 101 ms (incongruent)). To assess the effects of trial numbers on rank order consistency, different numbers of trials were sampled from these distributions (50, 100, 200, 400, 800, 1,600 and 3,200). In addition, Gaussian noise was added when sampling trials to simulate trial variability similar to that observed in Hedge et al.'s data. The noise sigma for reaction time and accuracy was optimized by minimizing the sum of squared errors of ICCs, ensuring alignment with the variability in the original data.

We then examined the impact of the number of trials and the number of participants on rank order consistency between the true mean and the apparent mean. The true mean for a participant was the value directly sampled from the between-participant distribution, while the apparent mean was the mean of n trials (n = 50, 100, 200, 400, 800, 1,600 and 3,200) sampled from the within-participant distribution, with added random noise. The number of simulated participants was also varied (50, 100, 200, 300, 400, 500, 1,000, 2,000 and 4,000). Finally, we calculated the absolute agreement across k measurements (ICC(A, k))[108] to assess the rank order consistency.

Simulation 3b—correlation between congruency effect and incongruent trial performance. Using the same simulation method described above with Hedge et al.'s (2018) data[32], we also examined the correlation between the congruency effect and performance on incongruent trials (Extended Data Fig. 2; reaction time and per cent error ((1 − accuracy) × 100); both measures on incongruent trials were expected to positively correlated with congruency effect). For this analysis, 500 participants were simulated, and for each participant, 3,200 trials were sampled to calculate the mean. These amounts were expected to give highly precise individual estimates. The correlation was calculated with Kendall's rank correlation coefficient and ICC(A, k).

Simulation 3c—effects of trial number on drift–diffusion modelling of inhibitory control. Using the same simulation method as simulation 3a, we examined how trial number affects the reliability of drift–diffusion modelling parameters (Supplementary Fig. 12). Due to its simplicity, ease of implementation and suitability for relatively sparse datasets, we performed EZ-diffusion modelling[62]. We simulated 100 participants and calculated the cross-participant reliability of the EZ-diffusion modelling parameters: drift rate, boundary separation and non-decision time. The key manipulation was to systematically increase the number of trials (50, 100, 200, 400, 800, 1,600 and 3,200). Sampling was done twice for each participant to assess test–retest reliability, using ICC(A, k). We compared these results with the reliability of the congruency effect reaction time and congruency effect accuracy.

Simulation 3d—effects of trial number on factor analyses of inhibitory control. We conducted CFA using simulated data based on Hedge et al.'s[32] flanker and Stroop task data (Extended Data Fig. 9). Similar to EZ-diffusion modelling, the goal was to observe how the number of trials per participant affects the reliability of the factor analysis. As a preliminary analysis, we ran P-technique factor analysis using the EPIC dataset to investigate whether the flanker task, prime–probe task and Stroop task share a latent factor. The best-fitting solution for six participants identified one factor, while for two participants, the best-fitting solution comprised two factors. Accordingly, we ran a CFA on the three tasks using a model containing one latent factor across varying numbers of trials per participant. We then simulated data using a method similar to that described for simulation 3a. While Hedge et al.'s dataset did not include a prime–probe task, simulations using only the flanker and Stroop tasks resulted in some simulated participants lacking a shared latent factor. To address this, we simulated prime–probe task data based on a combination of Hedge et al.'s Stroop task and the EPIC dataset's prime–probe task: the group mean and standard deviation were derived from the EPIC dataset, while individual distributions were based on Hedge et al.'s Stroop task data.

In addition to trial number, we manipulated cross-task correlation (at levels of 0.1, 0.4, 0.6, 0.8 and 1) by constructing three-dimensional probability distributions of the congruency effects across the three tasks. For the EZ-diffusion modelling simulation (simulation 3c), correlated samples for congruent and incongruent trials were generated separately by constructing a multivariate probability density that reflects the linear correlation between congruent and incongruent trials. By contrast, for this CFA simulation, a congruency effect value was simulated from the three-dimensional probability density, which captures the cross-task correlations among the three tasks. The CFA across the three tasks of 100 simulated participants was repeated twice to compute the test–retest reliability, and this process was repeated 100 times to obtain each participant's mean scores. The resulting reliability of each task's congruency effect and factor score was then plotted.

Simulation 4—comparing the traditional frequentist and the Bayesian approaches. We used WinBUGS and R packages (R2OpenBUGS and lme4) to compare the frequentist (non-hierarchical) and the Bayesian hierarchical modelling estimates of the congruency effect across several different conditions (Extended Data Fig. 10). Twenty-five replications of reaction time data were simulated per condition, based on flanker task data from the work of Robinson and Steyvers[49]. For the different conditions, we manipulated the number of trials (50, 100 and 500) and the ratio of within-participant variance to between-participant variance (5, 10, 20 and 40). These two factors were fully crossed to produce 12 total conditions. The number of simulated participants was fixed at 100. After the 25 datasets were simulated, we conducted multilevel modelling, using the lme4 package, to obtain unbiased estimates of between-participant variability in the congruency effect and trial-level variability in reaction time within participants. Next, using WinBUGS, Bayesian estimates of individual-level congruency effects were obtained. The unbiased estimates from multilevel modelling were used for the variance terms in the priors. For the frequentist non-hierarchical approach, individual-level congruency effects were estimated by simply calculating mean reaction time separately for congruent and incongruent trials, then taking the difference in these mean values for each participant. We evaluated the precision of the estimates with the mean absolute deviation between the generated (true) congruency effect and the estimated congruency effect values.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available via OSF at https://osf.io/jk9nb (https://doi.org/10.17605/OSF.IO/JK9NB).

Code availability

The code for the experimental tasks and data analyses conducted in this study is available via GitHub at https://github.com/GrattonLab/LeeSmith_EPIC.

References

1. Anderson, M. C. & Green, C. Suppressing unwanted memories by executive control. *Nature* **410**, 366–369 (2001).
2. Cook, P. B. & McReynolds, J. S. Lateral inhibition in the inner retina is important for spatial tuning of ganglion cells. *Nat. Neurosci.* **1**, 714–719 (1998).
3. Gratton, G., Cooper, P., Fabiani, M., Carter, C. S. & Karayanidis, F. Dynamics of cognitive control: theoretical bases, paradigms, and a view for the future. *Psychophysiology* **55**, e13016 (2018).
4. Mayr, U. Inhibition of action rules. *Psychon. Bull. Rev.* **9**, 93–99 (2002).
5. Müller, H. J. & Mühlenen, A. V. Probing distractor inhibition in visual search: Inhibition of return. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 1591–1605 (2000).
6. Munoz, D. P. & Everling, S. Look away: the anti-saccade task and the voluntary control of eye movement. *Nat. Rev. Neurosci.* **5**, 218–228 (2004).
7. Tipper, S. P. Does negative priming reflect inhibitory mechanisms? A review and integration of conflicting views. *Q. J. Exp. Psychol. Sect. A* **54**, 321–343 (2001).
8. Verbruggen, F. & Logan, G. D. Models of response inhibition in the stop-signal and stop-change paradigms. *Neurosci. Biobehav. Rev.* **33**, 647–661 (2009).
9. Darowski, E. S., Helder, E., Zacks, R. T., Hasher, L. & Hambrick, D. Z. Age-related differences in cognition: the role of distraction control. *Neuropsychology* **22**, 638–644 (2008).
10. Jaekel, J., Eryigit-Madzwamuse, S. & Wolke, D. Preterm toddlers' inhibitory control abilities predict attention regulation and academic achievement at age 8 years. *J. Pediatr.* **169**, 87–92 (2016).
11. Oberle, E. & Schonert-Reichl, K. A. Relations among peer acceptance, inhibitory control, and math achievement in early adolescence. *J. Appl. Dev. Psychol.* **34**, 45–51 (2013).
12. Abramovitch, A., Abramowitz, J. S. & Mittelman, A. The neuropsychology of adult obsessive–compulsive disorder: a meta-analysis. *Clin. Psychol. Rev.* **33**, 1163–1171 (2013).
13. Chamberlain, S. R., Blackwell, A. D., Fineberg, N. A., Robbins, T. W. & Sahakian, B. J. The neuropsychology of obsessive compulsive disorder: the importance of failures in cognitive and behavioural inhibition as candidate endophenotypic markers. *Neurosci. Biobehav. Rev.* **29**, 399–419 (2005).
14. Laurenson, C. et al. Cognitive control and schizophrenia: the greatest reliability of the Stroop task. *Psychiatry Res.* **227**, 10–16 (2015).
15. Westerhausen, R., Kompus, K. & Hugdahl, K. Impaired cognitive inhibition in schizophrenia: a meta-analysis of the Stroop interference effect. *Schizophr. Res.* **133**, 172–181 (2011).
16. Mullane, J. C., Corkum, P. V., Klein, R. M. & McLaughlin, E. Interference control in children with and without ADHD: a systematic review of flanker and Simon task performance. *Child Neuropsychol.* **15**, 321–342 (2009).
17. Casey, B. J. et al. The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
18. Eriksen, B. A. & Eriksen, C. W. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Percept. Psychophys.* **16**, 143–149 (1974).
19. Stroop, J. R. Studies of interference in serial verbal reactions. *J. Exp. Psychol.* **18**, 643–662 (1935).
20. Simon, J. R. & Rudell, A. P. Auditory S–R compatibility: the effect of an irrelevant cue on information processing. *J. Appl. Psychol.* **51**, 300–304 (1967).
21. Ebersole, C. R. et al. Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82 (2016).
22. Feldman, J. L. & Freitas, A. L. An investigation of the reliability and self-regulatory correlates of conflict adaptation. *Exp. Psychol.* **63**, 237–247 (2016).
23. MacLeod, C. M. Half a century of research on the Stroop effect: an integrative review. *Psychol. Bull.* **109**, 163–203 (1991).
24. Rey-Mermet, A., Gade, M. & Oberauer, K. Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**, 501–526 (2018).
25. Von Bastian, C. C. et al. Advancing the understanding of individual differences in attentional control: theoretical, methodological, and analytical considerations. *OSF* https://doi.org/10.31234/osf.io/x3b9k (2020).
26. Forstmann, B. U., Van Den Wildenberg, W. P. M. & Ridderinkhof, K. R. Neural mechanisms, temporal dynamics, and individual differences in interference control. *J. Cogn. Neurosci.* **20**, 1854–1865 (2008).
27. Janssens, C., De Loof, E., Boehler, C. N., Pourtois, G. & Verguts, T. Occipital alpha power reveals fast attentional inhibition of incongruent distractors. *Psychophysiology* **55**, e13011 (2018).
28. Mennes, M. et al. Linking inter-individual differences in neural activation and behavior to intrinsic brain dynamics. *NeuroImage* **54**, 2950–2959 (2011).
29. Wessel, J. R. An adaptive orienting theory of error processing. *Psychophysiology* **55**, e13041 (2018).
30. Zorowitz, S. & Niv, Y. Improving the reliability of cognitive task measures: a narrative review. *Biol. Psychiatry Cogn. Neurosci. Neuroimag.* **8**, 789–797 (2023).

31. Draheim, C., Mashburn, C. A., Martin, J. D. & Engle, R. W. Reaction time in differential and developmental research: a review and commentary on the problems and alternatives. *Psychol. Bull.* **145**, 508–535 (2019).
32. Hedge, C., Powell, G. & Sumner, P. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* **50**, 1166–1186 (2018).
33. Rouder, J. N. & Haaf, J. M. A psychometrics of individual differences in experimental tasks. *Psychon. Bull. Rev.* **26**, 452–467 (2019).
34. Dresler, T. et al. Reliability of the emotional Stroop task: an investigation of patients with panic disorder. *J. Psychiatr. Res.* **46**, 1243–1248 (2012).
35. Wilson, K. M. et al. Investigating the psychometric properties of the Suicide Stroop Task. *Psychol. Assess.* **31**, 1052–1061 (2019).
36. Eisenberg, I. W. et al. Uncovering the structure of self-regulation through data-driven ontology discovery. *Nat. Commun.* **10**, 2319 (2019).
37. Draheim, C., Pak, R., Draheim, A. A. & Engle, R. W. The role of attention control in complex real-world tasks. *Psychon. Bull. Rev.* **29**, 1143–1197 (2022).
38. Friedman, N. P. & Miyake, A. Unity and diversity of executive functions: individual differences as a window on cognitive structure. *Cortex* **86**, 186–204 (2017).
39. Rouder, J. N., Kumar, A. & Haaf, J. M. Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychon. Bull. Rev.* **30**, 2049–2066 (2023).
40. Elliott, M. L. et al. What is the test–retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychol. Sci.* **31**, 792–806 (2020).
41. Gell, M. et al. How measurement noise limits the accuracy of brain–behaviour predictions. *Nat. Commun.* **15**, 10678 (2024).
42. Nikolaidis, A. et al. Suboptimal phenotypic reliability impedes reproducible human neuroscience. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.22.501193 (2022).
43. Kong, R. et al. Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. *Cereb. Cortex* **29**, 2533–2551 (2019).
44. Pronk, T., Hirst, R. J., Wiers, R. W. & Murre, J. M. J. Can we measure individual differences in cognitive measures reliably via smartphones? A comparison of the flanker effect across device types and samples. *Behav. Res. Methods* **55**, 1641–1652 (2022).
45. Kucina, T. et al. Calibration of cognitive tests to address the reliability paradox for decision-conflict tasks. *Nat. Commun.* **14**, 2234 (2023).
46. White, C. N., Ratcliff, R. & Starns, J. J. Diffusion models of the flanker task: discrete versus gradual attentional selection. *Cogn. Psychol.* **63**, 210–238 (2011).
47. Friedman, N. P. et al. Individual differences in executive functions are almost entirely genetic in origin. *J. Exp. Psychol. Gen.* **137**, 201–225 (2008).
48. Rouder, J. N. & Mehrvarz, M. Hierarchical-model insights for planning and interpreting individual-difference studies of cognitive abilities. *Curr. Dir. Psychol. Sci.* **33**, 128–135 (2024).
49. Robinson, M. M. & Steyvers, M. Linking computational models of two core tasks of cognitive control. *Psychol. Rev.* **130**, 71–101 (2023).
50. Cronbach, L. J. & Furby, L. How we should measure 'change': or should we? *Psychol. Bull.* **74**, 68–80 (1970).
51. May, K. & Hittner, J. B. On the relation between power and reliability of difference scores. *Percept. Mot. Skills* **97**, 905–908 (2003).
52. Gershon, R. C. et al. NIH toolbox for assessment of neurological and behavioral function. *Neurology* **80**, S2–6 (2013).
53. Kim, S. & Cho, Y. S. Congruency sequence effect without feature integration and contingency learning. *Acta Psychol.* **149**, 60–68 (2014).
54. Notebaert, W., Gevers, W., Verbruggen, F. & Liefooghe, B. Top-down and bottom-up sequential modulations of congruency effects. *Psychon. Bull. Rev.* **13**, 112–117 (2006).
55. Weissman, D. H., Grant, L. D. & Jones, M. The congruency sequence effect in a modified prime-probe task indexes response-general control. *J. Exp. Psychol. Hum. Percept. Perform.* **46**, 1387–1396 (2020).
56. Ooi, L. Q. R. et al. Longer scans boost prediction and cut costs in brain-wide association studies. Preprint at *bioRxiv* https://doi.org/10.1101/2024.02.16.580448 (2024).
57. Weintraub, S. et al. The cognition battery of the NIH Toolbox for assessment of neurological and behavioral function: validation in an adult sample. *J. Int. Neuropsychol. Soc.* **20**, 567–578 (2014).
58. Baker, D. H. et al. Power contours: optimising sample size and precision in experimental psychology and human neuroscience. *Psychol. Methods* **26**, 295–314 (2021).
59. Davis-Stober, C. P., Dana, J. & Rouder, J. N. Estimation accuracy in the psychological sciences. *PLoS ONE* **13**, e0207239 (2018).
60. Weigard, A., Clark, D. A. & Sripada, C. Cognitive efficiency beats top-down control as a reliable individual difference dimension relevant to self-control. *Cognition* **215**, 104818 (2021).
61. Miyake, A. et al. The unity and diversity of executive functions and their contributions to complex 'frontal lobe' tasks: a latent variable analysis. *Cogn. Psychol.* **41**, 49–100 (2000).
62. Wagenmakers, E.-J., Van Der Maas, H. L. J. & Grasman, R. P. P. P. An EZ-diffusion model for response time and accuracy. *Psychon. Bull. Rev.* **14**, 3–22 (2007).
63. Haines, N. et al. A tutorial on using generative models to advance psychological science: Lessons from the reliability paradox. *Psychol. Methods* https://doi.org/10.1037/met0000674 (2025).
64. Whitehead, P. S., Brewer, G. A. & Blais, C. Are cognitive control processes reliable? *J. Exp. Psychol. Learn. Mem. Cogn.* **45**, 765–778 (2019).
65. Burgoyne, A. P., Tsukahara, J. S., Mashburn, C. A., Pak, R. & Engle, R. W. Nature and measurement of attention control. *J. Exp. Psychol. Gen.* **152**, 2369–2402 (2023).
66. Liesefeld, H. R. & Janczyk, M. Combining speed and accuracy to control for speed–accuracy trade-offs(?). *Behav. Res. Methods* **51**, 40–60 (2019).
67. Scarpina, F. & Tagini, S. The Stroop color and word test. *Front. Psychol.* **8**, (2017).
68. Vandierendonck, A. A comparison of methods to combine speed and accuracy measures of performance: a rejoinder on the binning procedure. *Behav. Res. Methods* **49**, 653–673 (2017).
69. Moretti, L., Koch, I., Hornjak, R. & Von Bastian, C. C. Quality over quantity: Focusing on high-conflict trials to improve the reliability and validity of attentional control measures. *J. Exp. Psychol. Learn. Mem. Cogn.* https://doi.org/10.1037/xlm0001466 (2025).
70. Matzke, D. et al. Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra Psychol.* **3**, 25 (2017).
71. Gordon, E. M. et al. Precision functional mapping of individual human brains. *Neuron* **95**, 791–807 (2017).
72. Gratton, C., Nelson, S. M. & Gordon, E. M. Brain–behavior correlations: two paths toward reliability. *Neuron* **110**, 1446–1449 (2022).
73. Protopapas, A., Vlahou, E. L., Moirou, D. & Ziaka, L. Word reading practice reduces Stroop interference in children. *Acta Psychol.* **148**, 204–208 (2014).
74. Dulaney, C. L. & Rogers, W. A. Mechanisms underlying reduction in Stroop interference with practice for young and old adults. *J. Exp. Psychol. Learn. Mem. Cogn.* **20**, 470–484 (1994).

75. Davidson, D. J., Zacks, R. T. & Williams, C. C. Stroop interference, practice, and aging. *Aging Neuropsychol. Cogn.* **10**, 85–98 (2003).
76. Erb, C. D., Germine, L. & Hartshorne, J. K. Cognitive control across the lifespan: congruency effects reveal divergent developmental trajectories. *J. Exp. Psychol. Gen.* **152**, 3285–3291 (2023).
77. Lansbergen, M. M., Kenemans, J. L. & Van Engeland, H. Stroop interference and attention-deficit/hyperactivity disorder: a review and meta-analysis. *Neuropsychology* **21**, 251–262 (2007).
78. Lipszyc, J. & Schachar, R. Inhibitory control and psychopathology: a meta-analysis of studies using the stop signal task. *J. Int. Neuropsychol. Soc.* **16**, 1064–1076 (2010).
79. Cha, C. B., Najmi, S., Park, J. M., Finn, C. T. & Nock, M. K. Attentional bias toward suicide-related stimuli predicts suicidal behavior. *J. Abnorm. Psychol.* **119**, 616–622 (2010).
80. Verdejo-García, A. J., Perales, J. C. & Pérez-García, M. Cognitive impulsivity in cocaine and heroin polysubstance abusers. *Addict. Behav.* **32**, 950–966 (2007).
81. Van Mourik, R., Oosterlaan, J. & Sergeant, J. A. The Stroop revisited: a meta-analysis of interference control in AD/HD. *J. Child Psychol. Psychiatry* **46**, 150–165 (2005).
82. Paap, K. R., Anders-Jefferson, R., Zimiga, B., Mason, L. & Mikulinsky, R. Interference scores have inadequate concurrent and convergent validity: should we stop using the flanker, Simon, and spatial Stroop tasks? *Cogn. Res. Princ. Implic.* **5**, 7 (2020).
83. Verhaeghen, P. & De Meersman, L. Aging and the Stroop effect: a meta-analysis. *Psychol. Aging* **13**, 120–126 (1998).
84. Bookheimer, S. Y. et al. The Lifespan Human Connectome Project in Aging: an overview. *NeuroImage* **185**, 335–348 (2019).
85. He, N., Rolls, E. T., Zhao, W. & Guo, S. Predicting human inhibitory control from brain structural MRI. *Brain Imaging Behav.* **14**, 2148–2158 (2020).
86. Kruschwitz, J. D., Waller, L., Daedelow, L. S., Walter, H. & Veer, I. M. General, crystallized and fluid intelligence are not associated with functional global network efficiency: a replication study with the human connectome project 1200 data set. *NeuroImage* **171**, 323–331 (2018).
87. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
88. Kong, R. et al. Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cereb. Cortex* **31**, 4477–4500 (2021).
89. Kadlec, J. et al. A measure of reliability convergence to select and optimize cognitive tasks for individual differences research. *Commun. Psychol.* **2**, 64 (2024).
90. Lee, H. J., Dworetsky, A., Labora, N. & Gratton, C. Using precision approaches to improve brain-behavior prediction. *Trends Cogn. Sci.* **29**, 170–183 (2025).
91. Banich, M. T. Executive function: the search for an integrated account. *Curr. Dir. Psychol. Sci.* **18**, 89–94 (2009).
92. Friedman, N. P. & Miyake, A. The relations among inhibition and interference control functions: a latent-variable analysis. *J. Exp. Psychol. Gen.* **133**, 101–135 (2004).
93. Gallagher, M. W. & Brown, T. A. in *Handbook of Quantitative Methods for Educational Research* (ed. Teo, T.) 289–314 (SensePublishers, 2013); https://doi.org/10.1007/978-94-6209-404-8_14
94. Van Essen, D. C. et al. The WU-Minn Human Connectome Project: an overview. *NeuroImage* **80**, 62–79 (2013).
95. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
96. Turner, B. O., Paul, E. J., Miller, M. B. & Barbey, A. K. Small sample sizes reduce the replicability of task-based fMRI studies. *Commun. Biol.* **1**, 62 (2018).
97. Bijsterbosch, J. Piggybacking on big data. *Nat. Neurosci.* **25**, 682–683 (2022).
98. He, T. et al. Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nat. Neurosci.* **25**, 795–804 (2022).
99. Kornblum, S., Hasbroucq, T. & Osman, A. Dimensional overlap: cognitive basis for stimulus–response compatibility—a model and taxonomy. *Psychol. Rev.* **97**, 253–270 (1990).
100. Nee, D. E., Wager, T. D. & Jonides, J. Interference resolution: insights from a meta-analysis of neuroimaging tasks. *Cogn. Affect. Behav. Neurosci.* **7**, 1–17 (2007).
101. Hazeltine, E., Lightman, E., Schwarb, H. & Schumacher, E. H. The boundaries of sequential modulations: evidence for set-level control. *J. Exp. Psychol. Hum. Percept. Perform.* **37**, 1898–1914 (2011).
102. Weissman, D. H., Egner, T., Hawks, Z. & Link, J. The congruency sequence effect emerges when the distracter precedes the target. *Acta Psychol.* **156**, 8–21 (2015).
103. Gratton, G., Coles, M. G. H. & Donchin, E. Optimizing the use of information: strategic control of activation of responses. *J. Exp. Psychol. Gen.* **121**, 480–506 (1992).
104. Braem, S. et al. Measuring adaptive control in conflict tasks. *Trends Cogn. Sci.* **23**, 769–783 (2019).
105. Redick, T. S., Calvo, A., Gay, C. E. & Engle, R. W. Working memory capacity and go/no-go task performance: selective effects of updating, maintenance, and inhibition. *J. Exp. Psychol. Learn. Mem. Cogn.* **37**, 308–324 (2011).
106. Bechtold, B. *Violin Plots for MATLAB* (GitHub Project, 2016).
107. Johnson, N. L., Kotz, S. & Balakrishnan, N. *Continuous Univariate Distributions* vol. 2 (Wiley, 1995).
108. McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* **1**, 30–46 (1996).

Acknowledgements

This work was supported by funds from NSF CAREER 2305698 (C.G.), NIH R01MH118370 (C.G.), T32NS047987 (D.M.S.) and R01MH121509 (D.E.N.). D.M.S. and C.E.H. also thank The Therapeutic Cognitive Neuroscience Fund. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the Article. This work was completed using the high-performance computing resources at the Research Computing Center at Florida State University and the Quest high-performance computing facility at Northwestern University. We thank S. Petersen, G. Gratton and M. Fabiani for their valuable comments on earlier drafts of this paper.

Author contributions

D.M.S., M.D. and C.G. designed the study. D.M.S. and A.D. collected data. A.D. stored and transported data. H.J.L., D.M.S., C.E.H., B.T.K., D.E.N. and C.G. analysed and interpreted data. H.J.L. conducted formal analyses and visualized results. H.J.L. and C.G. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-025-02198-2.

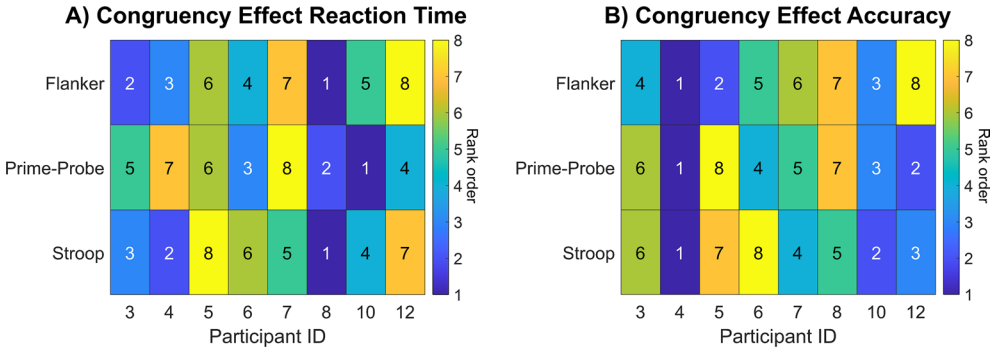
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02198-2.

Correspondence and requests for materials should be addressed to Hyejin J. Lee or Caterina Gratton.

Peer review information *Nature Human Behaviour* thanks Julia Haaf, Anna-Lena Schubert and the other, anonymous, reviewer(s) for their

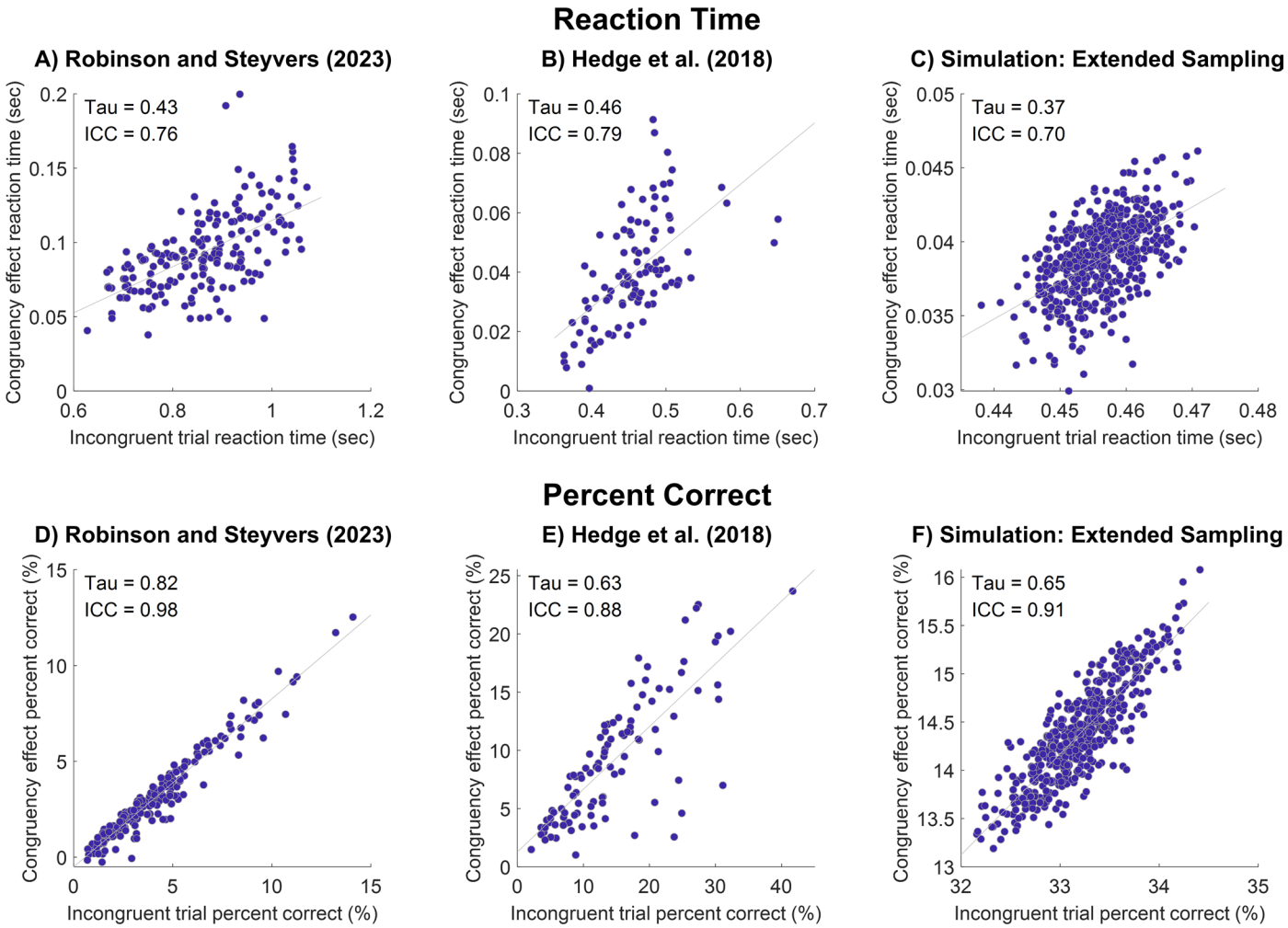
contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41562-025-02198-2



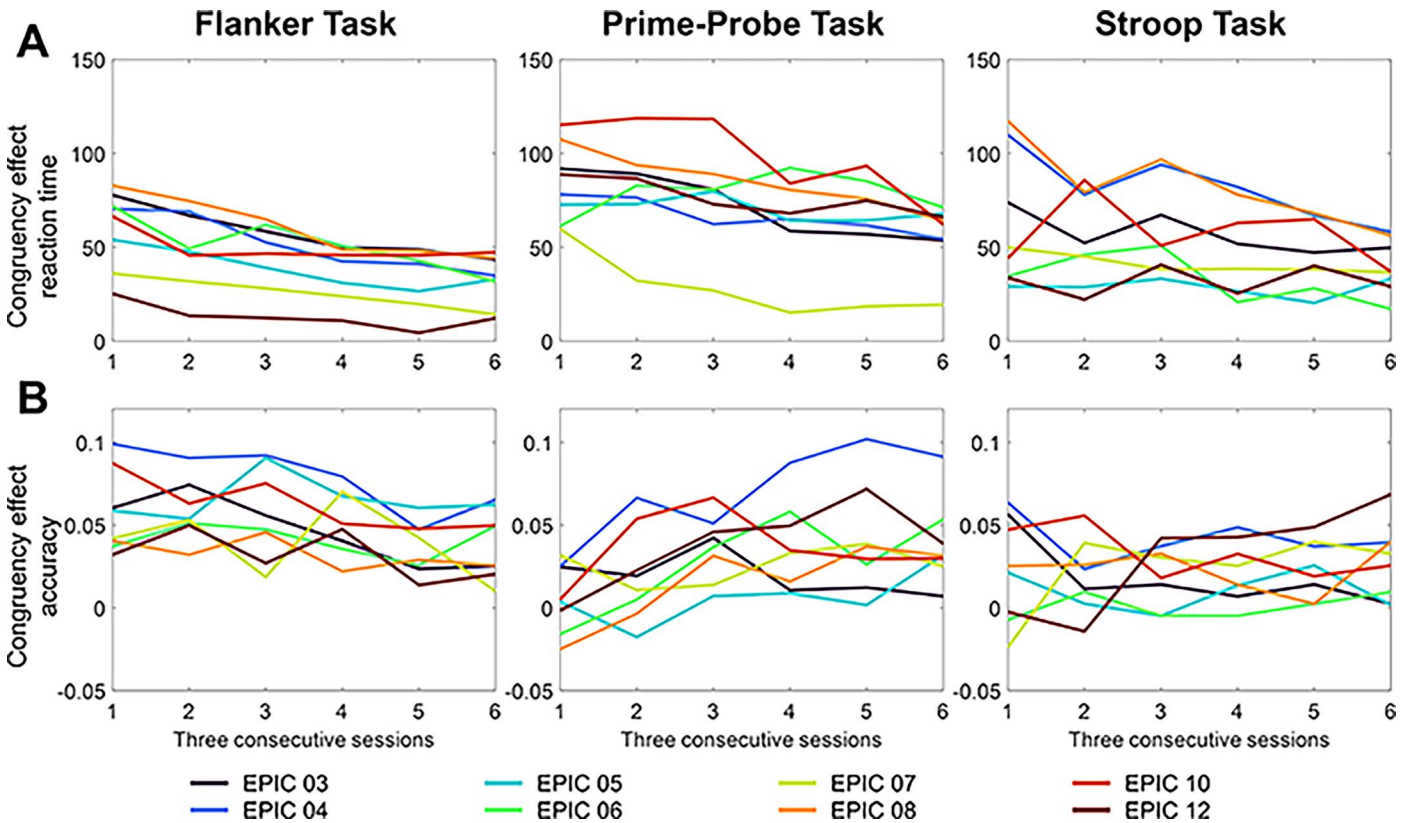
Extended Data Fig. 1 | Rank order matrices of the congruency effect. **a**) reaction time and **b**) accuracy from the EPIC data. The grand mean congruency effect for each participant across all sessions was calculated with the same exclusion criteria applied to remove outliers as for plotting Fig. 2. Then, we ranked for each task with participants exhibiting larger congruency effects ranked higher. Finally, we plotted these matrices to show the rank consistency across the three tasks. Despite our small sample, we observe notable consistency in the ranks, with the same or highly similar ranking across the three tasks (for example, EPIC 05 and 08 for reaction time, and EPIC 03, 04, 08 and 10 for accuracy). For both reaction time and accuracy, all participants show either the same rank or a difference of just one rank for at least two tasks, suggesting that rank orders across tasks can be consistent with extensive sampling.

https://doi.org/10.1038/s41562-025-02198-2



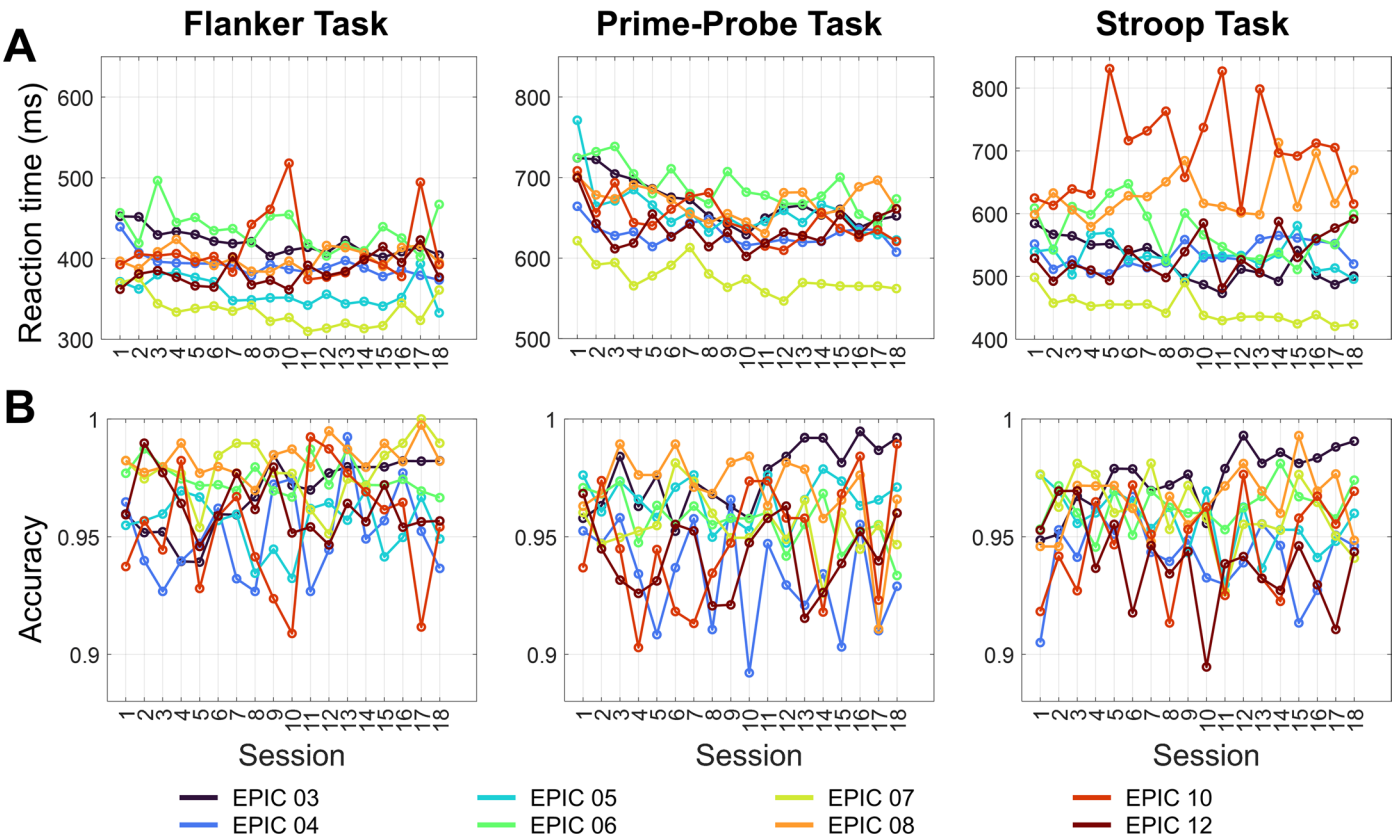
Extended Data Fig. 2 | Rank order consistency between congruency effect and incongruent trial performance. We used Robinson and Steyvers'[49] data (**a** and **d**), Hedge et al.'s[32] data (**b** and **e**) and simulated data with extended sampling (**c** and **f**). We simulated Hedge et al.'s empirical data by extending to 500 participants and 3,200 trials to resolve sampling variability (see *Methods* for details). To address the low reliability of the congruency effect, one proposed solution is to substitute it with performance on incongruent trials. However, this raises an important question of whether they measure the same construct. Our reaction time results demonstrate that, although they are correlated, the rank orders can still differ. Note that even with extended sampling, Kendall's $\tau = 0.37$ (ICC = 0.70). Interestingly, however, the rank order is more consistent for percent error results, yielding Kendall's $\tau = 0.65$ (ICC = 0.91) with extended sampling.

https://doi.org/10.1038/s41562-025-02198-2



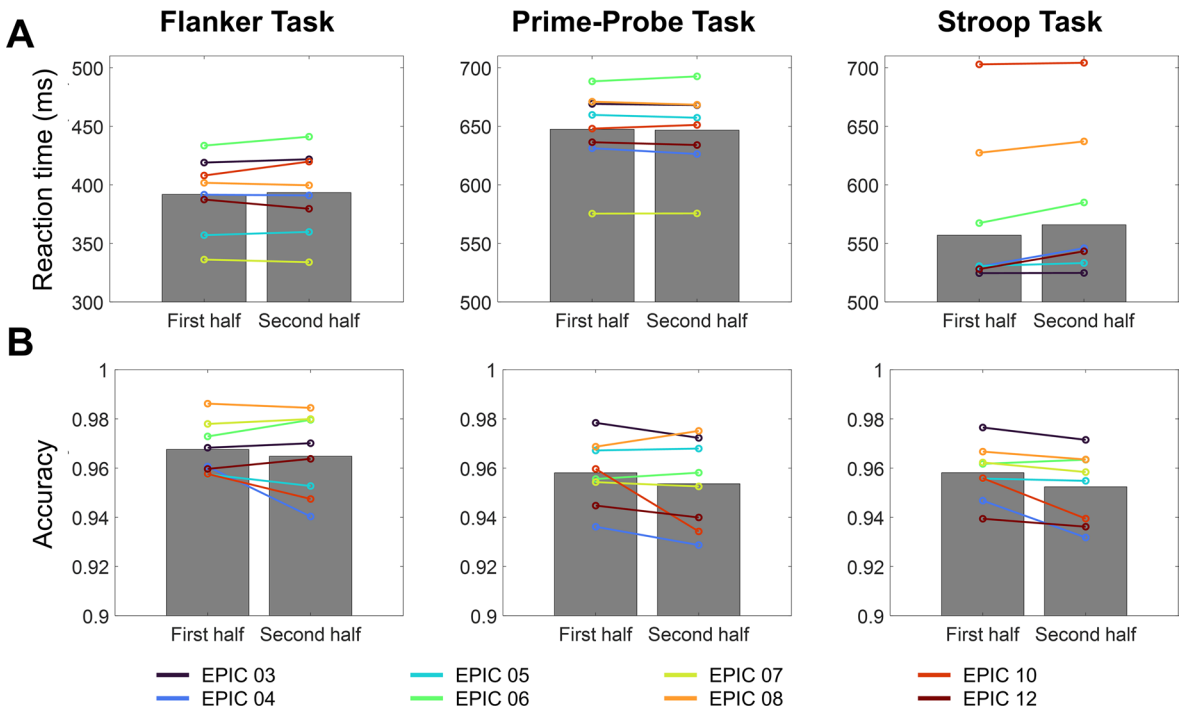
Extended Data Fig. 3 | Performance improvement in the congruency effect over time. a) reaction time and **b**) accuracy in the EPIC data. Each data point is the mean of three consecutive sessions, which yields approximately 1,000 trials when concatenated (1,200 for flanker, 1,152 for prime-probe and 864 for Stroop). This sampling was to minimize session-level variability (see Fig. 2 violin plots for variability across sessions). Notably, in reaction time data, congruency effects decrease in all three tasks, but the decrease is most prominent in the flanker task. Variation also exists across participants. It is possible that these effects observed over the course of sessions may be attributable to the time intervals between sessions, as Robinson and Steyvers[49] data do not show these effects. To address these effects, we regressed them out using a linear model (see Extended Data Fig. 6 for with and without linear regression). For accuracy, the trajectories seem relatively random, not displaying obvious linear trends as in reaction time data.

https://doi.org/10.1038/s41562-025-02198-2



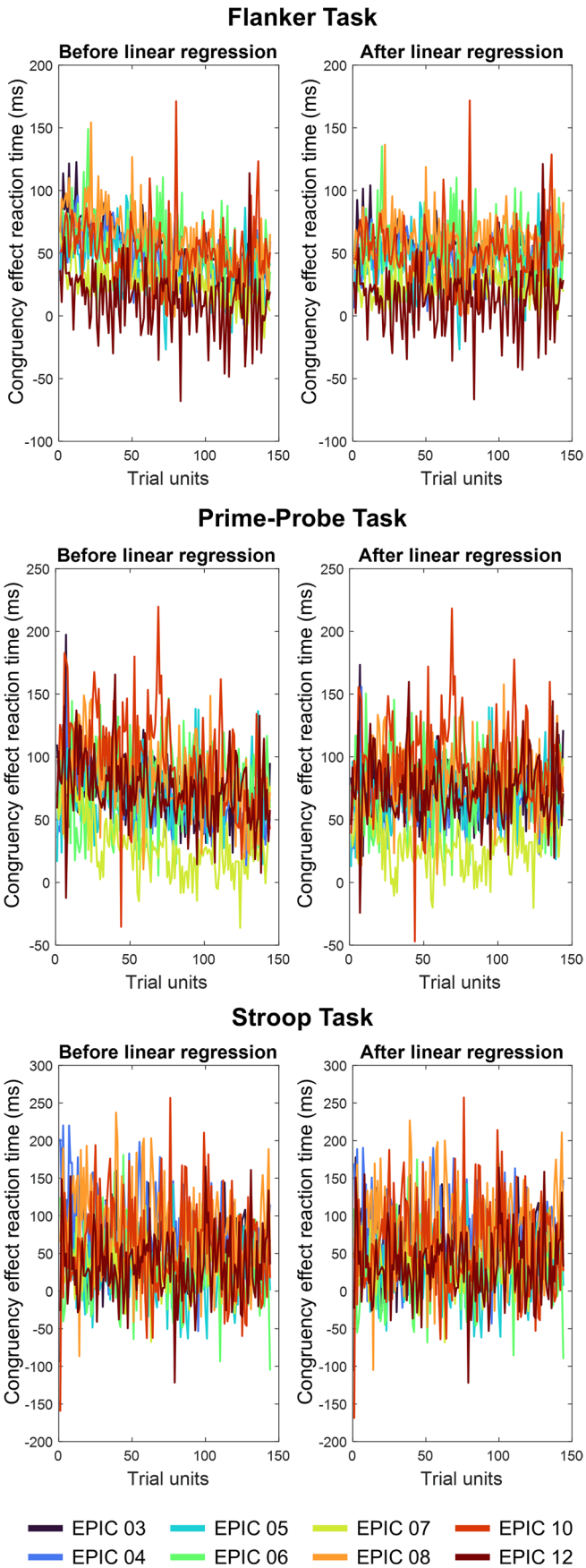
Extended Data Fig. 4 | Assessing performance impairment across sessions. This figure shows connected dot plots for performance on all trials (congruent and incongruent) combined, presented in terms of **a**) reaction time and **b**) accuracy across sessions of the EPIC data. For reaction time, incorrect trials and outlier (defined as those more than three standard deviations from the mean) were removed to calculate the mean for each session. We examined whether performance deteriorated in later sessions (that is, longer reaction times or poorer accuracy). The results show no systematic evidence of performance deterioration, as overall reaction time is generally consistent across sessions (except for EPIC 10). Overall accuracy shows some day-to-day variability, but around relatively small differences in close to ceiling performance (all participants showed greater than 89% accuracy in all sessions). Notably, while some participants show faster reaction times in later sessions (for example, EPIC 03 in all three tasks), this would more readily be interpreted as performance improvement due to practice, as accuracy is also higher for later sessions. In conclusion, although tested extensively across 18 sessions, we do not see substantial evidence for performance impairment across sessions in our dataset. This may be because our participants were a relatively homogeneous set who were highly motivated to participate in the study.

https://doi.org/10.1038/s41562-025-02198-2



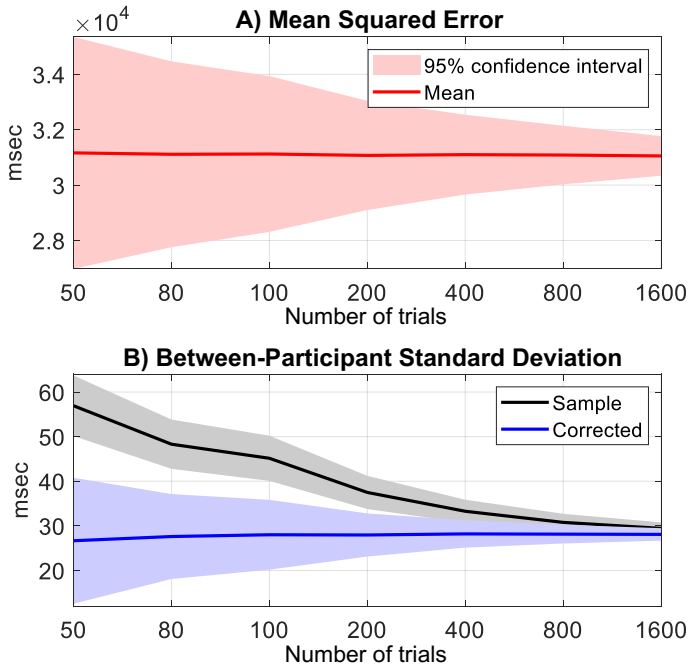
Extended Data Fig. 5 | Assessing performance impairment within sessions. This figure compares the first and second halves of the sessions of the EPIC data. Each session was divided into halves and the overall mean **a**) reaction time and **b**) accuracy were calculated for the 18 sessions. The goal was to examine whether any deterioration in performance within a session would occur, possibly due to factors such as boredom or fatigue, from testing approximately 400 trials per task (400 for flanker, 384 for prime-probe, and 432 for Stroop). Note that our participants performed two tasks in each session, so they were tested for about 800 trials in less than an hour. Each dotted line corresponds to one participant's mean, and the bar graphs show the group average of all participants. Except for the prime-probe task reaction time, overall, reaction time is higher, and accuracy is lower for the second half. However, statistical analyses (repeated measures ANOVA with session half as a variable separately conducted for each task and measure), show that only the difference between the two halves is significant for the Stroop task reaction time (with Bonferroni correction), $F(1, 7) = 13.25$, $p < 0.001$, $MSe = 23.92$, $\eta_p^2 = 0.65$. Note that the Stroop task had the most trials. For the flanker and prime-probe tasks, we did not observe significant performance degradation in the second halves of the sessions, $Fs < 1.72$. In sum, testing about 800 trials in a session, at least in our dataset, does not show significant performance degradation in the latter half of each session, although some (below threshold) impairment effects may be present. These results argue for not extending a single session to longer than the hour collected in this dataset.

https://doi.org/10.1038/s41562-025-02198-2



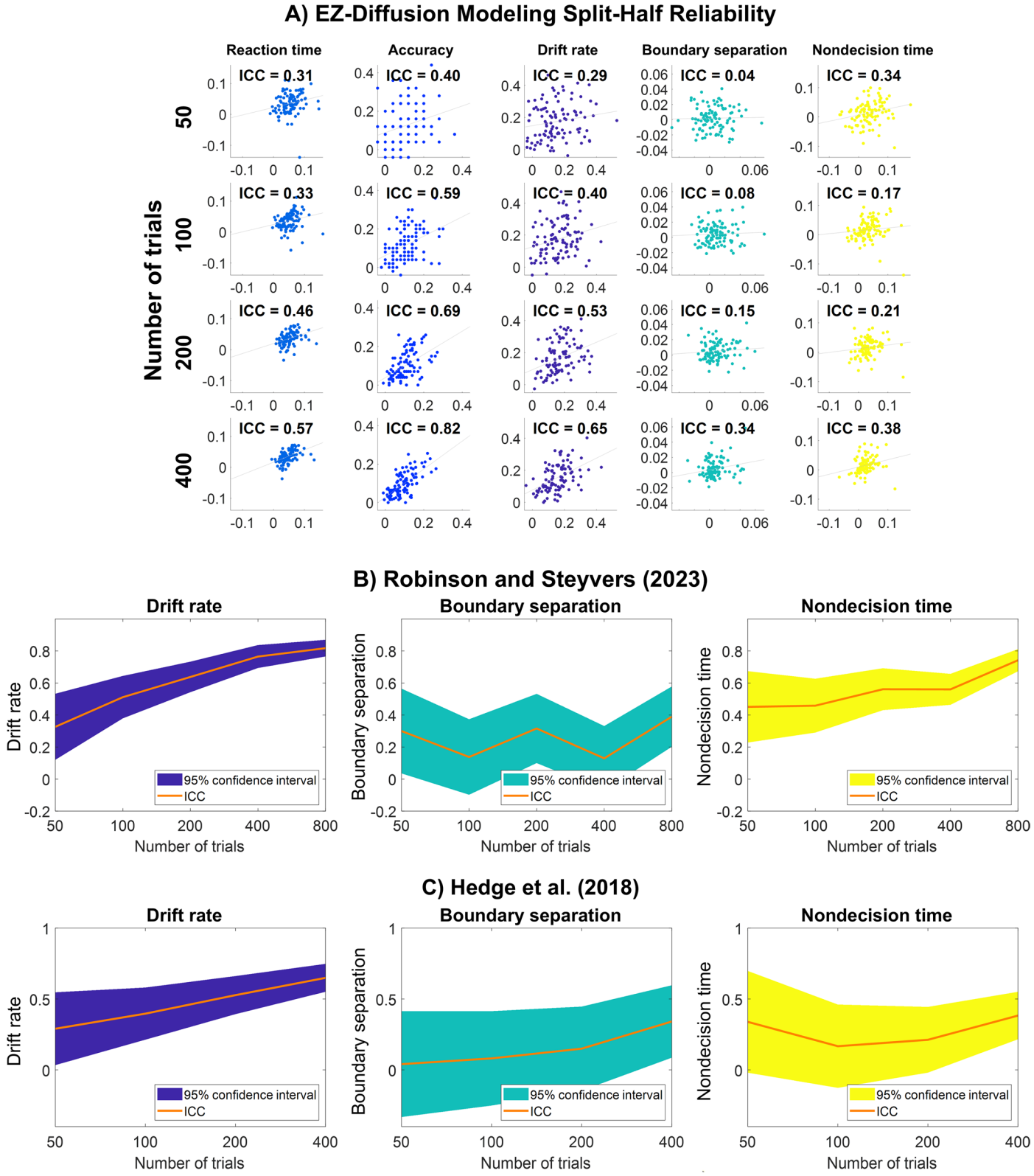
Extended Data Fig. 6 | Before and after linear regressions on the congruency effect. Related to Extended Data Fig. 3 that decreases in reaction time congruency effect are observed in the EPIC data, we regressed these effects with a linear model before implementing our two methods to draw stability curves. As the improvement effects are most prominent in the flanker task, the difference between before and after regression is also most noticeable for the flanker task. Notably, all participants exhibited congruency effects throughout the extent of data collection even when regressing out the improvement effects.

https://doi.org/10.1038/s41562-025-02198-2



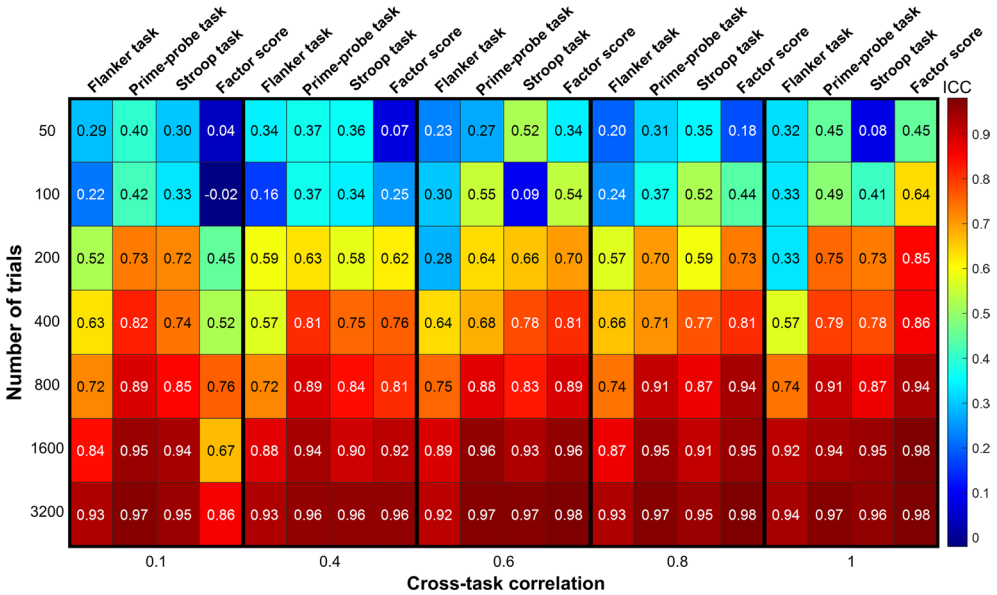
Extended Data Fig. 7 | Bias can be corrected but the imprecision may still be high without sufficient trials. a) Trial noise in congruency effects estimated with mean squared error as a function of number of trials. **b**) Sample versus corrected (by accounting for trial noise in the estimate) between-participant standard deviation. For both plots, the line indicates the mean of 1,000 simulations using Robinson and Steyvers'[49] data, and the shaded error bar is the 95% confidence interval. The goal here is to examine whether correcting between-participant standard deviation can be an effective strategy to get stable results when the number of trials is limited. We showed in Fig. 4 that the inflation of sample between-participant standard deviation can be rectified with sufficient trial sampling above 1,000 trials. Another effective way to correct the inflation is to separate trial noise from the sample between-participant variability[39]. Sample between-participant variance takes the following equation, $\frac{2\sigma^2}{T} + \sigma_d^2$, where σ_d^2 is true between-participant variance and $\frac{2\sigma^2}{T}$ is two times the within-participant variance (mean squared error) divided by trial number. We solved this equation for true between-participant variance (the 'corrected' value) and plotted it as a function of number of trials. We simulated data using Robinson and Steyvers' data parameters and sampled 25, 40, 50, 100, 200, 400 and 800 trials per condition, each for 1,000 times. We then plotted the mean and 95% confidence interval of the 1,000 simulations. Results show that correcting between-participant variability by accounting for trial noise in congruency effects effectively reduces bias/inflation and may be a promising approach that could be widely adopted. However, as the error bars show, the imprecision is still high with few trials. Thus, even with this correction method, a good estimate requires sufficient trials per participant. Note also that while this approach will help to reduce bias in estimates of between-participant variability, it does not give precise individual-level estimates of the congruency effect.

https://doi.org/10.1038/s41562-025-02198-2



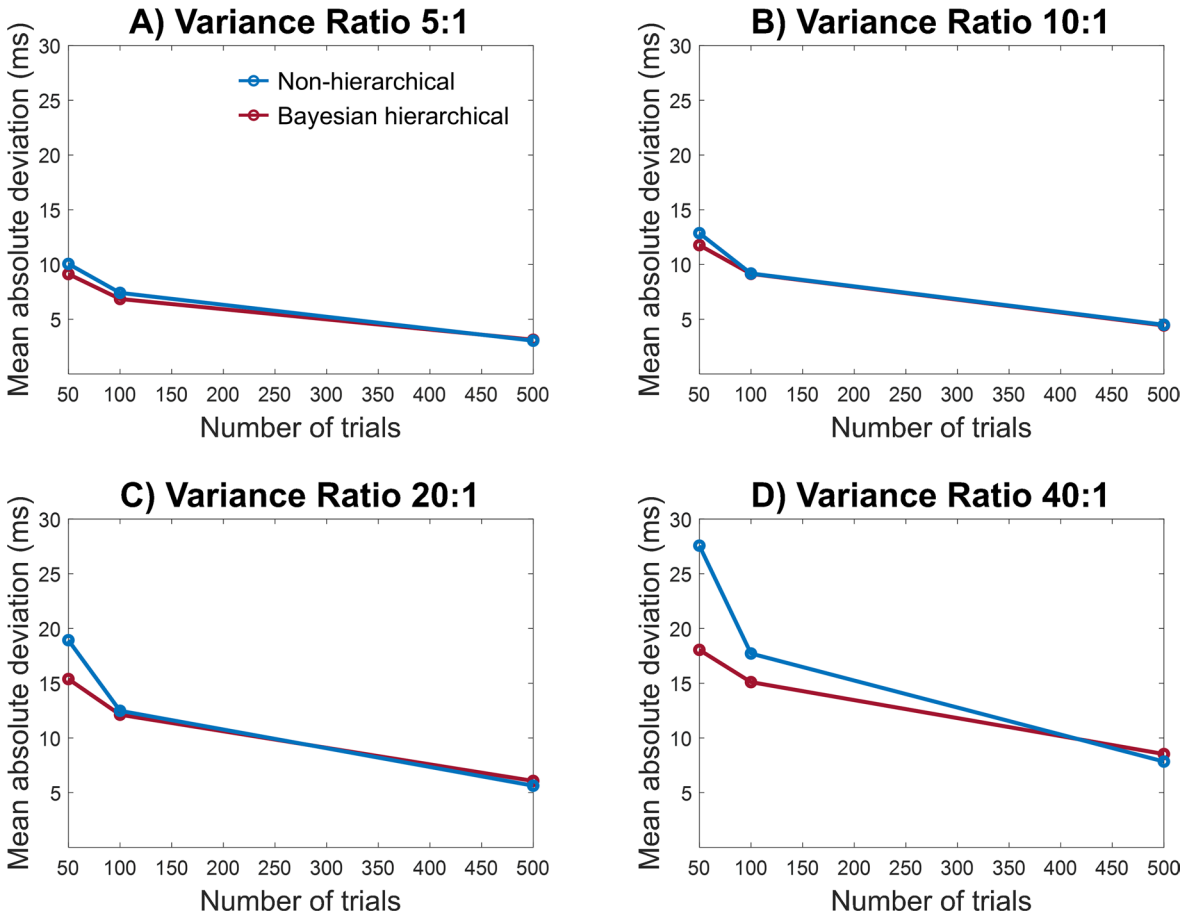
Extended Data Fig. 8 | Reliability of EZ-diffusion modelling with increased within-participant data. a) Split-half reliability of the EZ-diffusion modelling as a function of number of trials using Hedge et a.'s[32] empirical data. This plot corresponds to Fig. 8, which used Robinson and Steyvers'[49] data instead. The scatter plots and the ICCs of the congruency effect reaction time, accuracy, drift rate, nondecision time and boundary separation are shown. All results are difference scores between congruent and incongruent conditions. We increased the number of trials (50, 100, 200 and 400) sampled from the data to examine the effect of trial size on reliability. Results show that ICC increases with more trials, particularly for drift rate, consistent with Fig. 8. Below shows bootstrapped 95% confidence interval of the ICC using **b**) Robinson and Steyvers' data and **c**) Hedge et al.'s data to observe their precision across different number of trials. The orange lines indicate the ICC and the coloured shaded error bars are the 95% confidence interval of the ICC. The results show that ICC increases with more trials as well as its precision.

https://doi.org/10.1038/s41562-025-02198-2



Extended Data Fig. 9 | Reliability of confirmatory factor analysis across trial numbers. The data were simulated using the flanker task and Stroop task data from Hedge et al.[32], along with the EPIC prime-probe task data (see *Methods* for details on the simulation). The ICCs of the factor scores are plotted across increasing numbers of trials, compared to the ICCs of the individual task congruency effects. Based on P-technique factor analysis on the EPIC dataset's three tasks, we ran confirmatory factor analysis (CFA) on the simulated data, assuming one shared factor. We also manipulated the level of cross-task correlations when simulating the three task datasets (r = 0.1, 0.4, 0.6, 0.8 and 1). The results show that the test-retest reliability of factor scores improves with more trials, suggesting that CFA is also influenced by trial sampling size. Additionally, reliability critically depends on cross-task correlation; when cross-task correlation is high, the reliability of the factor score exceeds that of the individual task congruency effects.

https://doi.org/10.1038/s41562-025-02198-2



Extended Data Fig. 10 | Comparing frequentist non-hierarchical and Bayesian hierarchical approaches. To determine the extent to which a Bayesian hierarchical approach may or may not produce more precise congruency effect estimates than a frequentist non-hierarchical approach, we carried out a simulation, using Robinson and Steyvers[49] data, and compared parameter recovery between the two methods. Specifically, we evaluated the mean absolute deviation between the generated (true) congruency effect values and the recovered (estimated) congruency effect values. Mean absolute deviation values were then compared between the hierarchical Bayesian and non-hierarchical approaches. The manipulated factors in the simulation include the number of observed trials (50, 100, and 500) and the ratio of within-participant variance to between-participant variance, indexed as follows: A) 5, B) 10, C) 20 and D) 40. These factors were fully crossed, producing 12 total conditions. What is noteworthy is that there is an interaction effect with respect to the impact of trial number and within-participant variance on the hierarchical Bayesian improvement in mean absolute deviation. The hierarchical Bayesian approach offers a dramatic improvement when the number of trials is small and the within-participant variance is large. However, the precision of the hierarchical Bayesian and non-hierarchical estimates converge as trial number increases and within-participant variance decreases. By 500 trials, regardless of within-participant variance, there is virtually no difference in the precision of the estimates between the two methods. Note, additionally, that it requires at least 500 trials to obtain precision estimates in the target 4 to 9 ms range. Thus, if one's goal is to obtain estimates with a degree of precision in this target range, then a hierarchical Bayesian approach will not provide any additional benefit beyond a non-hierarchical approach. However, if resources are limited and one must settle for a non-optimal number of trials, a hierarchical Bayesian approach will provide utility.

nature portfolio

Corresponding author(s): Caterina Gratton
Hyejin J. Lee
Derek Smith

Last updated by author(s): 03/14/2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Task data were collected using custom codes written in Matlab (2018b) and Psychtoolbox 3.0.16 (available at https://github.com/GrattonLab/LeeSmith_EPIC). Survey data were collected with Qualtrics.
Data analysis	Data were analyzed using Matlab (2021b), WinBUGS (1.4), and R (4.3.2). All analyses codes are available at https://github.com/GrattonLab/LeeSmith_EPIC.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The main dataset collected for this project is available at https://osf.io/jk9nb.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We reported the sex of the participants in the manuscript. We did not conduct any sex- or gender-based analyses because our focus was on how inhibitory control measures may be stable or changing within an individual. Participants reported their sex in a Qualtric survey. The disaggregated sex information is available in the source data on the OSF platform (https://osf.io/jk9nb).
Reporting on race, ethnicity, or other socially relevant groupings	We collected data on the race and ethnicity of the participants, but did not use these variables, as our focus was on examining how task measures remain consistent or fluctuate within individuals.
Population characteristics	See above.
Recruitment	All participants were students at Northwestern University, which may introduce biases due to the recruitment of young, healthy adults from the same region with similar educational backgrounds. However, we demonstrate in the manuscript that our findings are generalizable, as they were replicated in more heterogeneous public datasets, including hundreds of participants from a broader age range (Robinson & Steyvers, 2023; Hedge et al., 2018). Potential participants from the Northwestern University Psychology Department laboratories were approached by study team members and invited to take part in a multi-week study on cognitive control. References Robinson, M. M., & Steyvers, M. (2023). Linking computational models of two core tasks of cognitive control. Psychological review, 130(1), 71-101. 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.
Ethics oversight	The institutional review board of Northwestern University

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study examined the number of trials needed to obtain maximally precise congruency effects. Multiple datasets and simulations show that extensive within-subject sampling is needed to achieve stable estimates of congruency effects.
Research sample	The participants in our main dataset were undergraduate and graduate students from Northwestern University (mean age = 25 years; age range: 18-30; five females, four males). This sample was accessible at the time the study was initiated during the COVID-19 pandemic. While this sample may not be fully representative of the general population due to its composition of young, healthy adults, it was selected for its ability to fully engage in our extensive testing sessions. To enhance the generalizability of our findings, we also utilized two publicly available datasets: Robinson and Steyvers (2023) and Hedge et al. (2018). Robinson and Steyvers collected online data from 485 participants via the Lumosity platform (mean age = 58; 66% female, 29% male; the rest did not report gender). Hedge et al. collected in-person data from 112 participants (mean age = 20.05; 15 males). References Robinson, M. M., & Steyvers, M. (2023). Linking computational models of two core tasks of cognitive control. Psychological review, 130(1), 71-101. 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.
Sampling strategy	We did not conduct a formal power analysis for this exploratory study. We were inspired in part by the Midnight Scan Club dataset (Gordon et al., 2017). This is a dataset comprised of only 10 subjects with a high level of within-subject data. The aim for the EPIC dataset was also to collect as much data per task from our participants as reasonably possible, making it a deep as opposed to a big dataset. The dataset was based on a sample of convenience and is comprised of graduate and undergraduate students. Due to the small sample size and the convenience sampling we also employed two large public datasets (Robinson & Steyvers, 2023; Hedge et al., 2018) to replicate the findings from our dataset. References

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., ... & Dosenbach, N. U. (2017). Precision functional mapping of individual human brains. Neuron, 95(4), 791-807.
Robinson, M. M., & Steyvers, M. (2023). Linking computational models of two core tasks of cognitive control. Psychological review, 130(1), 71-101.
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.

Data collection

Participants were either tested in the lab or at home, where they took the lab computer with them. During the lab sessions, only the participant and the researcher were present in the testing room. Participants working from home were instructed to work alone in a quiet environment. They were seated approximately 60 cm from an LCD monitor, which had a resolution of 1440x900 pixels and a refresh rate of 60 Hz. All experiments were programmed using MATLAB (2018b) and Psychtoolbox (3.0.16), and responses were collected via a standard computer keyboard. The researcher was not blinded to the study hypotheses, but the study was exploratory in nature and did not have firm hypotheses; the primary goal was to determine the number of trials needed to obtain maximally precise estimates of the congruency effect. Since the study did not involve experimental conditions (e.g., placebo vs. treatment), blinding of the researcher to such conditions was not relevant.

Timing

The start date of the first participant's data collection was 2020/5/18 and the end date of the last participant's data collection was 2022/3/8.

Data exclusions

We collected data from nine participants, but one was excluded due to an issue related to key release, producing consecutive error trials in later sessions.

We also used two public datasets to replicate findings from the main dataset. Robinson and Steyvers' (2023) data originally included 495 participants. The following excerpt from our manuscript's supplementary methods explains the participant exclusion procedure for this dataset:
"We excluded participants who had below 70% overall accuracy in either experimental conditions (congruent or incongruent) or who had 0% accuracy in any session, resulting in a total of 448 participants. We further selected participants with more than 2,500 correctly responded trials, resulting in 185 participants for our final analyses."
Details regarding the data collection procedures for Robinson and Steyvers' (2023) dataset can be found at the following link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10257386/

Hedge et al.'s (2018) dataset originally had 112 participants. The quoted text from our manuscript's supplementary methods explains the participant exclusion procedure for this dataset as follows:
"Five participants with only one session were excluded. ... Six participants, who had below 70% accuracy in either session, were excluded from analyses of both tasks, resulting in 101 participants."
Details regarding the data collection procedures for Hedge et al.'s (2018) dataset can be found at the following link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5990556/

References
Robinson, M. M., & Steyvers, M. (2023). Linking computational models of two core tasks of cognitive control. Psychological review, 130(1), 71-101.
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.

Non-participation

No participants dropped out of the study

Randomization

Participants were not placed into experimental groups

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.