You Could do Better Tomorrow:

Nonverbal reasoning fluctuates across days, working memory is stable

Michael E. Aristodemou[1], Nicholas Judd[1], Torkel Klingberg[2], and Rogier A. Kievit[1]

[1] *Cognitive Neuroscience Department, Donders Institute for Brain, Cognition and Behavior,Radboud University Medical Center, Nijmegen, the Netherlands*

[2] *Department of Neuroscience, Karolinska Institute, Stockholm, Sweden*

This paper has not been peer reviewed

Author note

Correspondence concerning this article should be addressed to Michael E. Aristodemou, M.Sc., Donders Center for Medical Neurosciences, Radboud University Medical Center, 6500 GL Nijmegen, the Netherlands.

Contact: michael.aristodemou@radboudumc.nl

Abstract

On some days we feel like we are not performing at our best. However, whether these experiences align with substantive differences in cognitive performance has not been studied systematically. We analyse dense time-series data of children's performance on nonverbal reasoning (n = 459) and visuospatial working memory (n = 4150) tasks using dynamic structural equation models to describe their pattern of instability across trials and days. Our model comparison confers domain-specific results, with children showing evidence for fluctuations in their nonverbal-reasoning speed from day-to-day, but stability in working-memory performance. The size of trial-to-trial fluctuations was weakly correlated with day-to-day fluctuations, suggesting distinct mechanisms across timescales. We show that day-to-day fluctuations in cognitive performance are more than folk intuition, argue that their neglect is problematic for translational and epistemic reasons, and demonstrate how a better understanding of cognitive performance as a dynamic phenomenon can improve cognitive assessment and theory construction.

Songs, sayings, and colloquialisms capture the shared experience of humans having ‘good days’ and ‘bad days’ – Ups and downs relative to some baseline on a given mental phenomenon of interest. In stark contrast, psychological constructs such as intelligence and personality are often considered hallmark examples of traits: Relatively stable features of people not expected to fluctuate meaningfully. But what if this assumption does not hold – If a person’s average score on a given day is substantially different from their score the day before or after? The widespread practice of single-occasion testing acts as a prominent example of the ramifications. Day-to-day instability in cognitive performance would mean that tests coinciding with either good or bad days could mis-stratify children, leading to lifelong consequences. This issue was first flagged by Raymond Cattell who, in 1966, stated that the neglect of performance fluctuations in intelligence testing would be ’morally wrong’ (p.357). Thus, taking good and bad days seriously could have considerable ethical and translational implications. These include re-examining performance thresholds for remedial teaching, incorporating estimates of daily variability into a more personalized assessment of a given child, and reforming the practice of standardized testing.

The existence of multiple mechanisms that vary from day-to-day and are, a priori, likely to affect cognitive performance suggests that the notion of good days and bad days is not just folk intuition. These include psychological mechanisms such as day-to-day fluctuations in sleep (Fang et al., 2021; Könen et al., 2015), mood (Kouros and El-Sheikh, 2015), fatigue (Hamilton et al., 2023), and self-regulation (Blume et al., 2022; Ludwig et al., 2016), all of which have been shown to affect cognitive performance (Blume et al., 2022; Chepenik et al., 2007; Kok, 2001; Könen et al., 2015; Lorist and Tops, 2003). Other plausible mechanisms are physiological. For instance, a considerable body of empirical evidence, informed by data from dedicated ‘health wearable’ tools have demonstrated the effect of patterns of fluctuation in, among others, heart-rate variability (Forte et al., 2019) respiration (Allen et al., 2023; Varga and Heck, 2017), and diurnal rhythms (Schmidt et al., 2007) on cognitive performance. We see evidence of daily fluctuations in all the above domains (Carr et al., 2018; Hunt et al., 1985; Kikuya et al., 2008; Robinson et al., 2014). This provides the impetus for us to expect that fluctuations in, at least some of, these domains bring about fluctuations in cognitive performance from day to day.

Taken together, cognitive fluctuations at the day-to-day level are plausible a priori, and understanding whether they occur is crucial for both principled and translational reasons. However, especially compared to cognitive performance, very little is known about cognitive fluctuations. In fact, only a few studies have described a pattern of day-to-day fluctuations (Dirk and Schmiedek, 2016; Galeano Weber et al., 2018; Galeano-Keiner et al., 2022; Rabbitt et al., 2001). Here we extend prior work by leveraging breakthroughs in quantitative approaches in a uniquely large and rich dataset. Using a flexible quantitative framework we use model comparison to formally test if day-to-day fluctuations are needed to explain performance in two widely studied cognitive constructs—visuospatial working memory (n = 4150) and nonverbal reasoning (n = 459).

Working memory is our first domain of choice. It has been shown to be highly predictive of academic performance (Alloway and Alloway, 2010; Friso-van den Bos et al., 2013; Gathercole et al., 2004), it has been implicated in multiple mental disorders (Grenard et al., 2008; Houben et al., 2011; Huang-Pollock et al., 2017; Nikolin et al., 2021), and is viewed as central to the structure of intelligence (Schmiedek et al., 2020). Our second cognitive domain of choice is nonverbal reasoning (or fluid intelligence). Nonverbal reasoning is ubiquitous in popular intelligence tests (e.g. WISC-V, Kaufman et al., 2015; Stanford-Binet intelligence scale 4th edition, Roid and Pomplun, 2012), it is highly correlated with standardized assessment scores (e.g. SAT; Frey and Detterman, 2004), and has been shown to have the largest impact on mathematics performance in a training study (Judd and Klingberg, 2021). Working memory and nonverbal reasoning research has shown that participants often showcase strategy shifts which form a plausible source of day-to-day fluctuations (Bobrowicz et al., 2024; Dunning and Holmes, 2014; Jastrzębski et al., 2018; Laurence and Macedo, 2023). Our hypotheses are as follows:

1. Day-to-day variability in *mean* visuospatial working-memory response speed and accuracy is needed to explain the pattern of children’s cognitive performance across multiple days.
2. Day-to-day variability in *mean* nonverbal reasoning speed is needed to explain the pattern of children’s cognitive performance across multiple days.
3. Individual differences in day-to-day variability and trial-to-trial variability will be positively associated

within each task. In other words, people who fluctuate more at the trial-to-trial level will also show more pronounced day-to-day fluctuations.

4. Individual differences in day-to-day variability will be positively associated across tasks – People who show more pronounced day-to-day effects in one task (visuospatial working memory) will also show pronounced day-to-day effects on the other (nonverbal reasoning).

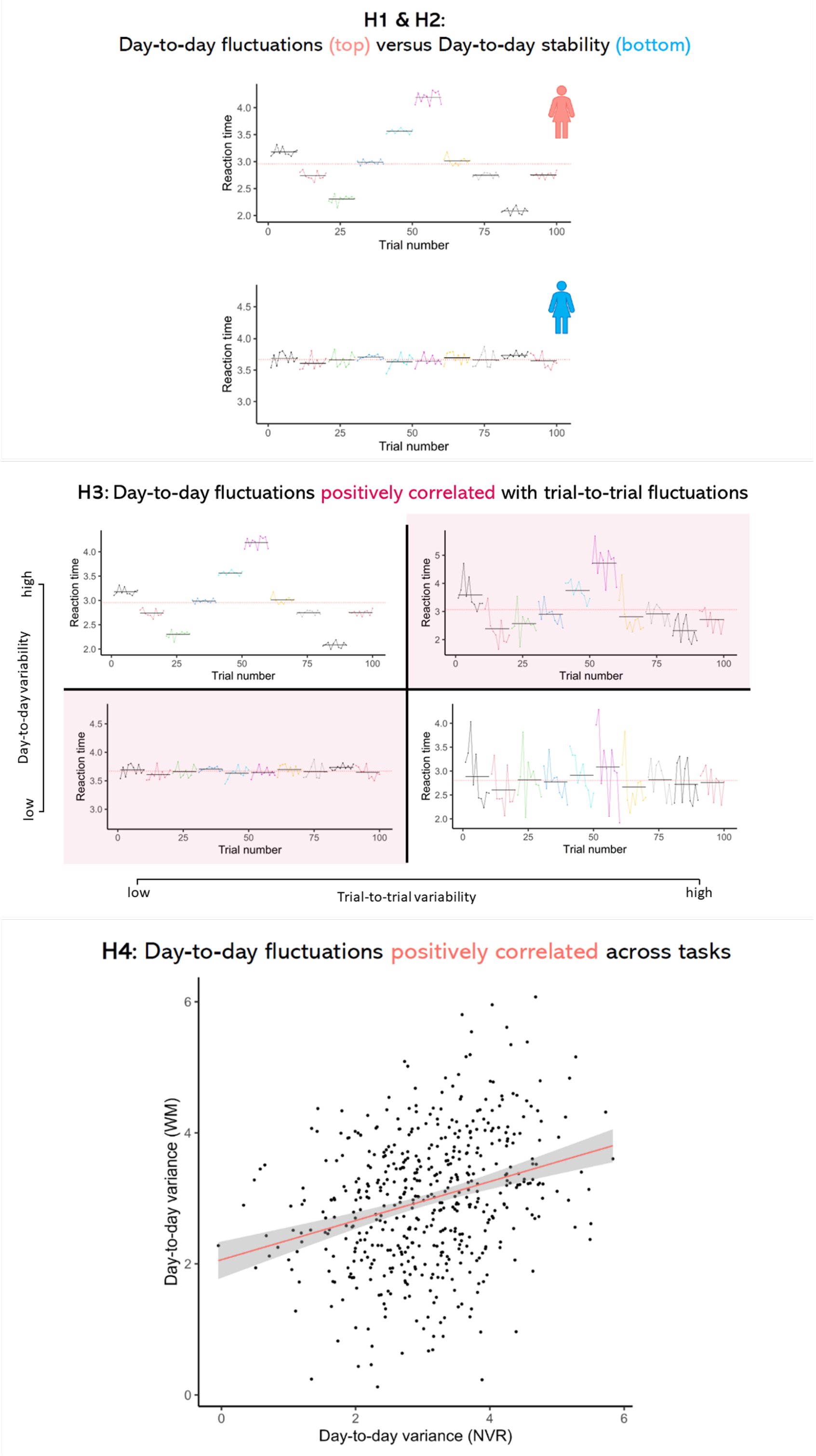

Figure 1. Visual illustration of hypotheses using simulated data. Hypotheses 1 and 2 (H1 & 2) are depicted in the top panel using two simulated examples that reflect the predictions of the models we are comparing. In the figure trials are color-coded to reflect different days. The mean of each day is depicted using solid black horizontal lines. The mean across all days is shown by the dashed red line running across all trials. For hypothesis 3 we visually depict our prediction using four quadrants. The shaded quadrants show patterns of data consistent with a positive correlation between day-to-day variance and trial-to-trial variance. Hypothesis 4 is depicted using a scatterplot where the association between individual differences in day-to-day variance across tasks is expected to be positive.

Methods

Participants

We analyzed data from two subsamples of children enrolled in the Vektor app. Vektor is an adaptive cognitive training app designed to improve children's mathematics performance through cognitive training (Judd and Klingberg, 2021). Children were voluntarily enrolled for Vektor training by their respective educators. The Vektor app includes multiple training plans to which it randomly allocates children once they log in. For our analyses, we included children who completed at least 20 days of visuospatial working memory training (n = 4150) or 20 days of nonverbal reasoning training (n = 459). We chose this cutoff to mirror the minimum day count of prior research that found evidence for reliable fluctuations from day-to-day Dirk and Schmiedek, 2016). Demographic characteristics of children such as sex, gender, and ethnicity were not logged through the app. Children were between five and eight years old at the time of assessment.

Procedure and materials

Working memory grid task

To assess visuospatial working memory performance, a working memory grid task was used (Figure 2b, bottom panel). The working memory grid task displays a 4x4 grid of dots. The dots light up in a sequence that needs to be recalled by the participant. The difficulty of the task adapts to the child's performance. Each time a participant correctly reproduces a sequence of dots two times in a row the difficulty increases by 1 level. If the participant incorrectly recalls the sequence once, the difficulty decreases by 1 level. The amount of items that can be recalled (i.e. level number) can be seen as a proxy for the child's working memory span. Children's performance is recorded using an ordinal measure of accuracy, using the number of sequential dots correctly recalled within a trial (i.e. partial-credit load scoring; Conway et al., 2005), and a continuous measure of response speed in milliseconds. Accuracy was treated as a continuous measure for all analyses since its range exceeds 5 values and its distribution is Gaussian (see Supplementary Materials, Figure SA1;

Rhemtulla et al., 2012).

Nonverbal reasoning task

To assess nonverbal reasoning (NVR) performance, children performed a task where they had to complete a sequence of spatial patterns by choosing the correct option from a variable number of alternatives (Figure 2b, top panel). The app modulates difficulty by increasing the complexity of stimuli in the sequence (i.e., increasing number of dots, colors, and shapes, Bergman Nutley et al., 2011; Judd and Klingberg, 2021). Children's performance was recorded using a continuous measure of response speed in milliseconds.

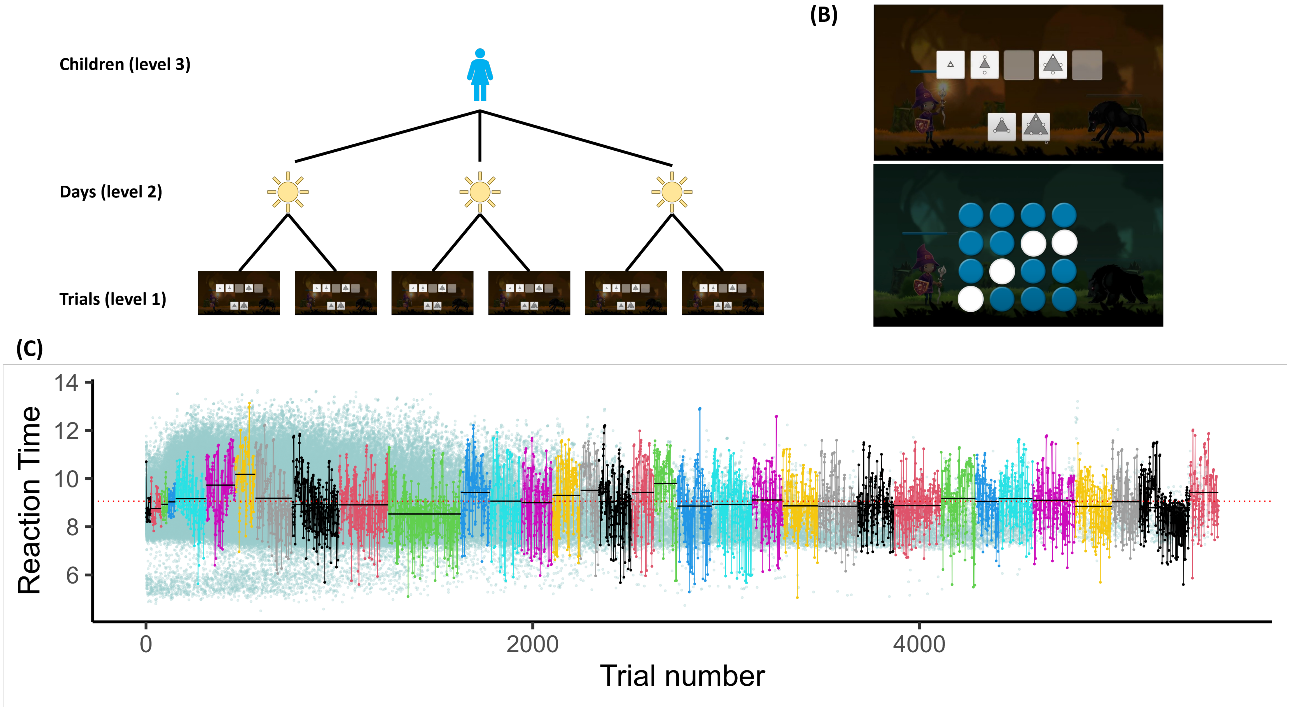

Figure 2. (a) Schematic representation of data nested data structure with trials nested within days, nested within children. (b) Screenshots of sequential nonverbal reasoning task (top panel) and visuospatial working-memory-grid task (bottom panel). (c) Raw data of a single child with performance during separate days color coded, superimposed on top of a scatterplot of light-blue dots showing the performance of all children in our sample. The child's mean performance during each day is marked by the black horizontal lines and their mean performance across all days is marked using a dashed red line.

Statistical analyses

Pre-processing

First, for the working memory task we divided the response time on each trial by accuracy, defined as the number of buttons correctly pressed in a sequence. Since longer sequences take longer to mechanically reproduce this correction should better approximate comparable response-times across trials. For the nonverbal reasoning task no correction was applied, since there was no relation between trial difficulty and artefact-driven response speed. Second, for both tasks, response times were log-transformed before any analyses to reduce skewness (Supplementary Materials, Figure SA1). The log-transformed response times were used for all analyses.

Summary of analyses

We used two complementary modeling approaches to test our hypotheses. First, we extended the Dynamic Structural Equation modeling (DSEM; Asparouhov et al., 2018; Lin et al., 2018; Rast and Ferrer, 2018) framework by building a three-level DSEM to model our nested data structure with trials (level 1), nested within days (level 2), nested within children (level 3; Figure 2a). To assess whether, on average, children in our population fluctuated from day-to-day in their cognitive performance (visuospatial working memory and nonverbal reasoning) we compared a two-level DSEM, that does not explicitly model variation at the day-to-day level, to our three-level DSEM. Testing a three-level structure in one step allows us to estimate the population day-to-day variance while taking the uncertainty of estimates into account and applying shrinkage. Second, we assessed how many children showed evidence for daily fluctuations via model selection at the level of each individual using models that estimate the parameters of dynamic structural equation models for each child separately. This allowed us to assess individual differences in day-to-day variance as a complement to our first analysis. The trade is that when we aggregate individual-level estimates in further analyses we do not take their uncertainty into account. Third, we assessed whether individual differences in day-to-day variability and trial-to-trial variability are correlated within each task. Fourth, we assessed whether individual differences in day-to-day variability are correlated across tasks.

Bayesian Dynamic Structural Equation models

Our goal was to test for the existence of day-to-day fluctuations against a simpler alternative (i.e. baseline model) which assumes that average performance is comparable across days. We specified a two-level DSEM as our baseline model. DSEM encompasses four trial-level sources of variation and individual differences therein that are frequently found in human time-series, and reflect substantively interesting cognitive processes in and of themselves. First, we have *mean performance* which captures a child's average performance across the entire time series (γ_{00i}). Second, we have *trial-to-trial variability* (σ_i). Since fluctuations at faster timescales need to be taken into account when modeling fluctuations at slower timescales (Schmiedek et al., 2013; Dirk & Schmiedek, 2016), we modelled trial-to-trial fluctuations which are the most well-studied type of instability with a growing mechanistic literature surrounding them (Li et al., 2001; Fagot et al., 2018; Garrett et al., 2023). Third, *inertia* (first-order autoregression) captures how well performance on a given trial predicts performance on the subsequent trial (β_{2i}). Temporal dependency amongst adjacent trials is a typical feature of human time-series (e.g., Solfo and Van Leeuwen, 2023; Wagenmakers et al., n.d.). Failing to account for *inertia* can inflate our estimates of daily fluctuations (de Haan-Rietdijk et al., 2016). Lastly, we have a *linear trend* ($\beta 1i$). Repeated exposure to a task can induce improvement (or worsening, through e.g. boredom) in performance over days. This also needs to be taken into account, lest we misinterpret growth for day-to-day fluctuations. We within-person mean centered the inertia parameter (de Haan-Rietdijk et al., 2016). Our three-level DSEM model, adds one parameter to the two-level DSEM which allows each child's mean performance to vary from day-to-day using random effects. This parameter is termed the *day-to-day variability* parameter (β_{0ij}) and was used to assess whether at the population level we find evidence of day-to-day fluctuations in mean cognitive performance via model selection. We within-person/within-day mean centred the inertia parameter in the three-level DSEM model. This allows us to estimate the mean of each day, instead of an intercept (de Haan-Rietdijk et al., 2016). We allowed individual differences in each one of the DSEM parameters, except for day-to-day variability, due to limitations in currently available tools. The mathematical specification of this model for subject *i*, at day *j*, on trial *k*, is a follows:

Location (mean) model:

Level 1 model:

$$y_{ijk} = \beta_{0ij} + \beta_{1i}Time_{ik} + \beta_{2i}y_{(k-1)i} + \epsilon_{ijk}$$

Level 2 model:

$$\beta_{0ij} = \gamma_{00i} + \upsilon_{0ij}$$

Level 3 model:

$$\gamma_{00i} = \delta_{00} + \nu_{00i}$$

$$\beta_{1i} = \delta_{01} + \nu_{01i}$$

$$\beta_{2i} = \delta_{02} + \nu_{02i}$$

Scale (variance) model:

$$log_{(}\sigma^2_{ijk}) = \omega_0 + \tau_i$$

Model estimation

We estimated our Bayesian models using the R package rstan (Carpenter et al., 2017). Rstan is a compiler for the probabilistic programming language Stan in C++. We also include code using the brms package for accessibility (Bürkner, 2017). The brms package is a front-end for the statistical computing language Stan and is written in the accessible style of lme4. The Stan models are estimated using No-U-turn sampling (NUTS), an extension of Hamiltonian Markov Chain Monte Carlo (MCMC) estimation (Hoffman and Walters, 2022). Due to computational limitations, we created random subsets of approximately 200 participants and ran our models for each subset separately. This includes 2 subsets for analyses using nonverbal reasoning speed and 20 subsets for analyses using visuospatial working memory performance. We found negligible deviations in parameter estimates across subsets (see Supplementary Materials B).

Prior specification

We used diffuse priors to mirror the priors used in Mplus with the exception of half-Cauchy priors for the scale parameters as suggested by Gelman, 2006 (see Supplementary materials C for priors).

Model diagnostics

To check the robustness of our Bayesian models we ran a series of model diagnostic tests based on the WAMBS guidelines (Aristodemou et al., 2022; Depaoli and van de Schoot, 2017; van de Schoot et al., 2021). For a detailed description of our diagnostic process and results please see Supplementary materials D.

Model comparison

Our primary aim was to assess whether day-to-day fluctuations were necessary to explain children's time series of cognitive performance. To do this, we ran two models: (1) the three-level DSEM and (2) the two-level DSEM. We used leave-one-out cross validation using the loo package in R to compare the prediction error between the two models. The model with the lowest prediction error was chosen as the best model.

Individual-level models: DSEM for each individual

To assess whether day-to-day variability is necessary to explain children's time series at an individual level we ran a two-level DSEM for each child separately. This can be seen as a frequentist version of the two-level DSEM applied to individual-level data with levels representing trials (level 1) nested within days (level 2). These models are an extension of the models used by Galeano-Weber et al., 2018 and contain the following parameters: (a) mean performance, which reflects the average performance over all trials (γ_{0j}); (b) linear trend, which captures the systematic change in performance over the course of the training plan ($\beta_1 Time_k$); (c) inertia, which captures the extent to which performance on a given trial (t) can be predicted by performance on the previous trial (t-1; $\beta_2 y_{(k-1)}$); (d) trial-to-trial variability, reflects the magnitude of a child's fluctuations from their mean performance from trial-to-trial (ϵ_{jk}); (e) day-to-day variability, reflects the magnitude of daily fluctuations in mean performance from a child's overall mean performance (υ_{0j}). We within-day mean center the inertia parameter. The mathematical specification of the individual-level model for day j on trial k is:

Level 1 model:

$$y_{jk} = \beta_{0j} + \beta_1 Time_k + \beta_2 y_{(k-1)} + \epsilon_{jk}$$

Level 2 model:

$$\beta_{0j} = \gamma_{00} + \upsilon_{0j}$$

Model comparison

First, we wanted to know if day-to-day variability was necessary to explain children's time series of cognitive performance for each child individually. We compared two models using two information criteria, which quantify the evidence in favor of each model while explicitly penalizing models for having more parameters, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). We abstained from reusing leave-one out cross validation to test the robustness of our model comparison across fit indices. The two individual-level (n = 1) models differed with respect to one parameter, the inclusion of day-to-day variability. Substantively, this amounts to one model predicting that a child's mean performance will fluctuate from day-to-day while the other model predicting stability in a child's mean performance across all days. For each child we conducted the model comparison and computed the difference in AIC and BIC between the two models.

Extracting individual difference estimates

After specifying our models we aggregated the parameter estimates from the individual-level models to obtain individual-difference estimates of these parameters. We used these estimates to assess the covariance between individual differences in day-to-day variability and trial-to-trial variability within tasks, and day-to-day variability between tasks.

Results

Task 1: Visuospatial working memory grid

Visuospatial working memory performance is stable from day-to-day

Bayesian Dynamic Structural Equation models

We did not find evidence in favor of day-to-day fluctuations in working memory performance in our population. Specifically, including a parameter which captured average day-to-day fluctuations in working memory speed (0.02 SD) and accuracy (0.22 SD) did not improve the predictive power of our baseline model. This was determined by comparing a three-level DSEM that included day-to-day variance with a baseline DSEM that did not, using leave-one-out cross validation (($WMspeed_{ELPD_{diff}} = -55.8, WMspeed_{SE_{diff}} = 25.2; WMaccuracy_{ELPD_{diff}} = -2202.1, WMaccuracy_{SE_{diff}} = 52.4$)). For all parameter estimates please refer to Supplementary Materials B.

Individual-level models: DSEM for each individual

The above result was mirrored in our individual-level model comparison. Sixty out of 4150 children (1.45%) manifested a substantial amount of day-to-day variability in working memory speed and 112 children fluctuated from day-to-day in their accuracy (2.7%), according to our information criteria (Figure 3).

Visuospatial working memory speed. Children's working memory response-speed fluctuated on average by 0.05 standard deviations (SD) from day-to-day ($mean_{daySD} = 0.04, SD_{daySD} = 0.05, range_{daySD} = 0.00 - 0.42$). These estimates are slightly larger in comparison to our Bayesian three-level model, since information across participants is not being used to inform individual-level estimates and the aggregation of these estimates ignores uncertainty. To put this effect size into perspective, on average, children's day-to-day fluctuations are approximately 29% of the magnitude of individual differences in children's working memory speed ($mean_{logRT} = 7.96, SD_{logRT} = 0.12$; Figure 4). Trial-to-trial fluctuations were on average

3.44 times greater than the between-subject variation in children's working memory speed with a standard deviation of 0.71 ($mean_{trialSD} = 0.43, SD_{trialSD} = 0.07, range_{trialSD} = 0.19 - 0.88$).

Visuospatial working memory accuracy. Children's working memory accuracy fluctuated on average by 0.37 SD from day-to-day ($mean_{daySD} = 0.37, SD_{daySD} = 0.14, range_{daySD} = 0.00 - 1.15$). The average day-to-day variability was approximately 60% of the magnitude of individual differences in children's working accuracy ($mean_{accuracy} = 2.86, SD_{accuracy} = 0.61, range_{accuracy} = 0.78 - 5.24$). The magnitude of average trial-to-trial fluctuations was 1.79 times greater than the magnitude of between-subject differences in working memory speed ($mean_{trialSD} = 1.09, SD_{trialSD} = 0.14, range_{trialSD} = 0.48 - 1.61$). For all the parameter estimates the Supplementary materials E. For a correlation matrix of all parameter estimates see Supplementary materials F.

Task 2: Nonverbal reasoning

Nonverbal reasoning speed fluctuates from day-to-day

Bayesian DSEM

In contrast with our findings for visuospatial working memory, our analysis of nonverbal reasoning speed suggested substantial evidence in favour of day-to-day differences in performance. Specifically, the average magnitude of day-to-day fluctuations (0.42 SD) in nonverbal reasoning speed was substantial. This was determined by comparing a three-level to a two-level DSEM using leave-one-out cross validation ($ELPD_{diff} = -55.8, SE_{diff} = 25.2$).

Individual-level models

Approximately half of the children in our sample showed substantial day-to-day fluctuations in nonverbal reasoning speed according to our information criteria (AIC 47%, BIC 41%; Figure 3).

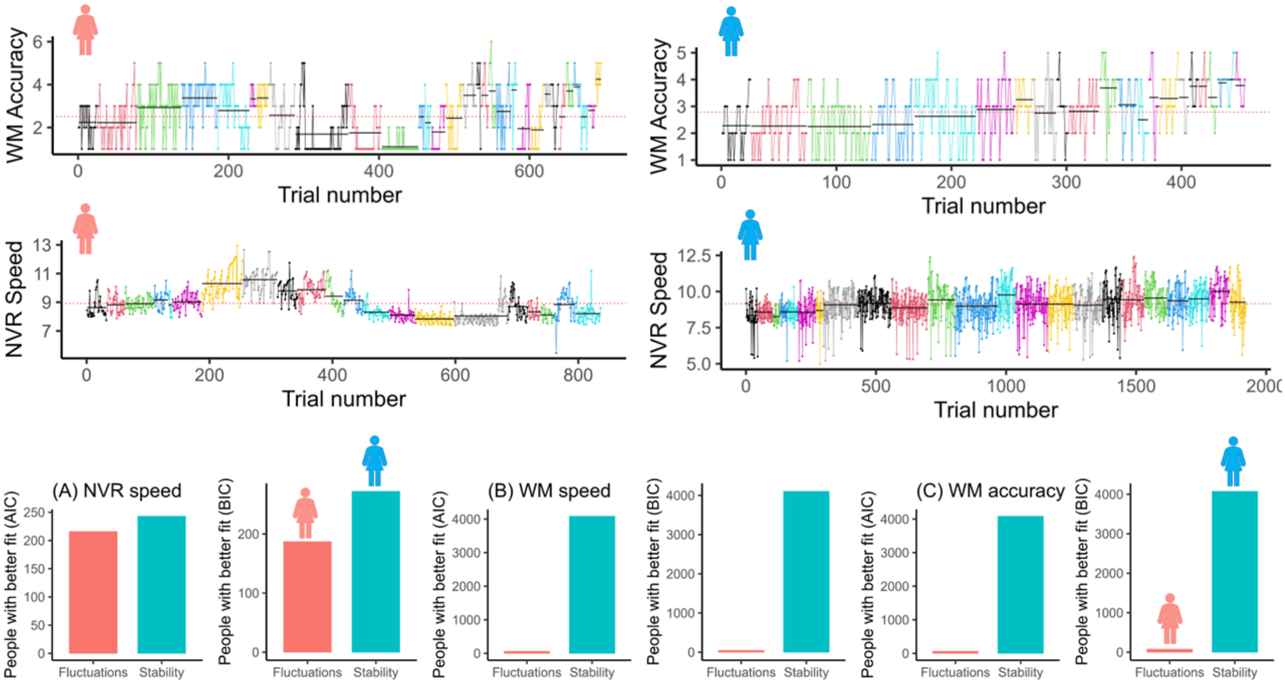

Figure 3. Barplots and lineplots of participants classified as having day-to-day fluctuations (red) or stable performance (blue). Lineplots illustrate raw data from participants in each classification. The bar plots show the absolute count of participants that have better fit for each model according to information criteria AIC and BIC in nonverbal reasoning speed (A) and working memory speed (B) and working memory accuracy (C). The models were compared for each individual separately and include a model that estimates day-to-day fluctuations in performance (red) and one that constraints day-to-day fluctuations to zero (blue)

Nonverbal reasoning speed. Children's nonverbal response speed fluctuated on average by 0.40 standard deviations (SD) from day-to-day ($mean_{daySD} = 0.40, SD = 0.10, range = 0.21 - 0.77$). Children's day-to-day fluctuations were 1.69 greater than the magnitude of individual differences in children's nonverbal reasoning speed ($mean_{logRT} = 8.93, SD = 0.23$; Figure 4). Trial-to-trial fluctuations were on average 2.93 times greater than the between-subject variation in children's working memory speed with a standard deviation of 0.69 ($mean_{trialSD} = 0.69, SD = 0.10, range = 0.42 - 1.10$). Children who fluctuated more from day-to-day also showed greater trial-to-trial fluctuations (r = 0.22). The moderate positive association indicates that individual differences across these two timescales are partially overlapping, but are largely driven by non-overlapping factors.

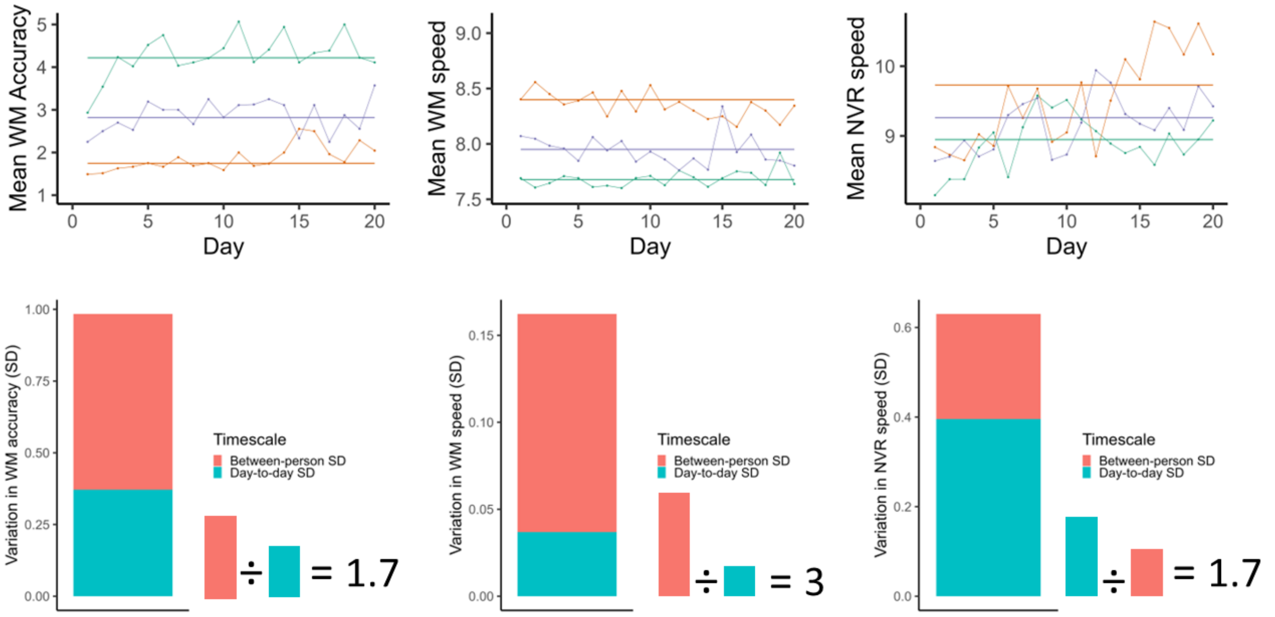

Figure 4. Stacked barplots showing the total amount of day-to-day variability and between-subject variance, along with the relative amount of variation that each component accounts for. Line plots show data from participants used to illustrate either stability of rank order from day-to-day or day-to-day fluctuations in rank-order.

Cross-domain correlations

Children with higher trial-to-trial variance in nonverbal reasoning also had greater trial-to-trial variance in working memory speed (r = 0.30) but lower trial-to-trial variance in accuracy (r = -0.13). Focusing on mean performance instead of variability, we find that children with slower mean nonverbal reasoning response speed also had slower visuospatial working memory speed (r = 0.25) and lower accuracy (r = -0.11). Complete information on correlations between all three outcomes across the two cognitive domains can be found in the Supplementary materials F.

Discussion

Children's cognitive performance fluctuates from day-to-day, but not in all cognitive domains. The size of day-to-day fluctuations in children's nonverbal reasoning speed varied per individual, with approximately half of the children in our sample showing substantial day-to-day variance. The average magnitude of fluctuations was 1.7 times larger than the between-subject variance in mean nonverbal-reasoning speed. Meaning that children's performance hierarchy in our sample is reshuffled across days. Trial-to-trial fluctuations were on average greater than day-to-day fluctuations. The two timescales showed a moderate positive correlation. This indicates some commonality or interaction in the forces shaping fluctuations across timescales. Although, most of the variance within timescales seems to be driven by distinct sources.

Intriguingly, and contrary to our hypotheses, children's visuospatial working memory performance was stable from day-to-day (accuracy and speed). This finding is at odds with three studies that reported evidence in favour of day-to-day fluctuations in working memory performance (Dirk and Schmiedek, 2016; Galeano Weber et al., 2018; Galeano-Keiner et al., 2022). Our study presents the most well-powered examination of fluctuations in working memory, controls for statistical properties of human time-series that can inflate estimates of day-to-day variability (e.g., temporal dependency; see de Haan-Rietdijk et al., 2016), used adaptive testing, and evaluation metrics that explicitly reward parsimony. Hence, we believe current evidence favors stability in children's day-to-day mean working memory performance. However, at a timescale of seconds working memory performance was volatile. Trial-to-trial fluctuations were on average 1.8 or 2.4 times greater than between-subject differences in mean performance, depending on our performance metric.

The domain-specificity of our findings suggests that the mechanisms driving day-to-day nonverbal reasoning fluctuations specifically influence the reasoning domain. This aligns with theories that view working memory as a distinct component which restricts our reasoning ability by constraining the complexity of the structural representations we can form (Chuderski et al., 2012; Hagemann et al., 2023; Shipstead et al., 2012; Oberauer, Süß, Wilhem, & Sander, 2007). Our results suggest the working memory part of this equation is constant. While day-to-day fluctuations capture instability in a unique component of nonverbal reasoning which allows us to understand how it is distinct from working memory performance. Since mechanisms with a broad effect on cognition, such as daily shifts in diurnal rhythms, require more assumptions to fit this scenario it may be productive to seek more targeted mechanisms (e.g. active strategy shifting; Laurence and Macedo, 2023). Shifts in strategies can explain some of the day-to-day variation in nonverbal reasoning speed, while leaving working memory performance unaffected. Individual differences in day-to-day fluctuations can in turn arise due to propensities for directed exploration which are tied to differences in prefrontal dopamine function, creating a link between physiological and behavioural mechanisms (Frank et al., 2009). It remains an open question why individual differences in directed exploration did not manifest as differences in day-to-day fluctuations in the working memory task which can also rely on strategy use. Future work combing dense behavioral testing with neurophysiological measurements can employ joint modeling techniques to test which latent processes give rise to day-to-day fluctuations in cognitive performance.

Differences in the propensity to shift strategies can also explain the overlap between children that have greater trial-to-trial variance and day-to-day variance in nonverbal reasoning performance. The positive and moderate association suggests that a propensity to shift strategies cannot exhaustively explain variation at both the trial and day-to-day levels. This also renders common explanations of trial-to-trial variance such as neural noise (Li et al., 2006; Voytek et al., 2015) dopamine binding potential (MacDonald et al., 2006), and inattention (Aristodemou et al., 2022; Cai et al., 2021) incomplete explanations of day-to-day fluctuations in nonverbal reasoning speed. The moderate correlation between the two timescales may also indicate that there are “good” and “bad” types of variation at both the trial-to-trial and day-to-day level. For instance, strategy shifts are arguably a “good” type of variability if they lead to the discovery of a better solution. In contrast, sources like sleeplessness are less likely to be beneficial. The weak positive correlation may then be a byproduct of some children showing a negative (or null) correlation between timescales, due to for example higher “bad” trial variance and lower “good” day-to-day variance. Clustering approaches based on patterns of covariation in time-series characteristics (Ntekouli et al., 2023) offer a promising method to identify individual differences in the processes underlying fluctuations across timescales.

Our findings suggest that cognitive phenotyping needs cognitive fluctuations. Fluctuations at both the trial and day-level were weakly associated with mean performance supporting their potential to provide unique information essential for the comprehensive assessment of cognitive performance (Judd et al., 2023). Trial-level fluctuations are already available in the data assessors routinely collect. Integrating this information would lead to a better understanding of children's performance at negligible cost. Moreover, as contemporary evidence challenges the static trait conceptualization of cognitive performance, we need to re-evaluate what the mean reflects. While there seems to be an anchor point around which fluctuations happen, this is not the same as our understanding of the mean as a static true ability. This opens the door for a new theoretical framework that does not necessary relegate the importance of the mean, but makes space for dynamic properties which could offer effective intervention targets (Epstein et al., 2011; Kofler et al., 2013) and the necessary constraints to our theories.

Our discussion should be qualified by our limitations. First, important demographics characteristics that were not logged through the Vektor application could serve as covariates in our analyses and inform the generalizability of our results. Second, current modeling techniques are improving but still come with limitations. Aggregating estimates from our single-case models disregards uncertainty in estimates. We reconciled this by modeling the three-level structure in a Bayesian framework with the trade of omitting individual differences in day-to-day variance. Thus, our modeling approaches complement each other. Future efforts to create three-level models that allow for simultaneous estimation of individual differences across all levels are becoming increasingly desirable as we move toward a dynamic understanding of psychological constructs. Third, to obtain a reliable estimate of day-to-day variance we selected subsamples of children with at least 20 days of training in visuospatial working memory and nonverbal reasoning tasks. This could induce a selection of children who vary systematically from those who could not or did not perform at least 20 days. Future studies with random assignment of children to a set amount of training days can address this issue.

The idea that cognition fluctuates from day-to-day has been around for decades, but rarely tested in a formalized, quantitative manner. We conducted the first principled test of day-to-day fluctuations in cognitive performance in a large cohort of 4150 children. We found substantial daily fluctuations in nonverbal reasoning speed and day-to-day stability in working memory performance. Fluctuations provide information that is unique from mean performance and allow for a more granular dissection of children's cognitive performance. This should encourage researchers to move away from identifying intelligence with a static snapshot of performance and embrace the signal within motion. The potential reward for those keen to explore dynamics in cognitive performance is a better understanding of children's cognitive capacities, fairer assessment strategies, and new cognitive interventions.

References

Allen, M., Varga, S., & Heck, D. H. (2023). Respiratory rhythms of the predictive mind. *Psychological Review*, *130*(4), 1066–1080. https://doi.org/10.1037/rev0000391

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Aristodemou, M., Rommelse, N., & Kievit, R. (2022, July 15). *Attentiveness modulates reaction-time variability: Findings from a population-based sample of 1032 children* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/j2n5w

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Bergman Nutley, S., Söderqvist, S., Bryde, S., Thorell, L. B., Humphreys, K., & Klingberg, T. (2011). Gains in fluid intelligence after training non-verbal reasoning in 4-year-old children: A controlled, randomized study [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-7687.2010.01022.x]. *Developmental Science*, *14*(3), 591–601. https://doi.org/10.1111/j.1467-7687.2010.01022.x

Blume, F., Irmer, A., Dirk, J., & Schmiedek, F. (2022). Day-to-day variation in students' academic success: The role of self-regulation, working memory, and achievement goals [_eprint: https://onlinelibrary.wiley.com/doi/pdf *Developmental Science*, *25*(6), e13301. https://doi.org/10.1111/desc.13301

Bobrowicz, K., Weber, A., & Greiff, S. (2024). The successful use of a search strategy improves with visuospatial working memory in 2- to 4.5-year-olds. *Journal of Experimental Child Psychology*, *238*, 105786. https://doi.org/10.1016/j.jecp.2023.105786

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Cai, W., Warren, S. L., Duberg, K., Pennington, B., Hinshaw, S. P., & Menon, V. (2021). Latent brain state dynamics distinguish behavioral variability, impaired decision-making, and inattention [Number: 9 Publisher: Nature Publishing Group]. *Molecular Psychiatry*, *26*(9), 4944–4957. https://doi.org/10.1038/s41380-021-01022-3

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*, 1. https://doi.org/10.18637/jss.v076.i01

Carr, O., Saunders, K. E. A., Tsanas, A., Bilderbeck, A. C., Palmius, N., Geddes, J. R., Foster, R., Goodwin, G. M., & De Vos, M. (2018). Variability in phase and amplitude of diurnal rhythms is related to variation of mood in bipolar and borderline personality disorder. *Scientific Reports*, *8*(1), 1649. https://doi.org/10.1038/s41598-018-19888-9

Chepenik, L. G., Cornew, L. A., & Farah, M. J. (2007). The influence of sad mood on cognition. *Emotion*, *7*(4), 802–811. https://doi.org/10.1037/1528-3542.7.4.802

Chuderski, A., Taraday, M., Nęcka, E., & Smoleń, T. (2012). Storage capacity explains fluid intelligence but executive control does not. *Intelligence*, *40*(3), 278–295. https://doi.org/10.1016/j.intell.2012.02.010

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

de Haan-Rietdijk, S., Kuppens, P., & Hamaker, E. L. (2016). What's in a day? a guide to decomposing the variance in intensive longitudinal data. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00891

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in bayesian statistics: The WAMBS-checklist. *Psychological Methods*, *22*(2), 240–261. https://doi.org/10.1037/met0000065

Dirk, J., & Schmiedek, F. (2016). Fluctuations in elementary school children's working memory performance in the school context. *Journal of Educational Psychology*, *108*(5), 722–739. https://doi.org/10.1037/edu0000076

Dunning, D. L., & Holmes, J. (2014). Does working memory training promote the use of strategies on untrained working memory tasks? *Memory & Cognition*, *42*(6), 854–862. https://doi.org/10.3758/s13421-014-0410-5

Epstein, J. N., Brinkman, W. B., Froehlich, T., Langberg, J. M., Narad, M. E., Antonini, T. N., Shiels, K., Simon, J. O., & Altaye, M. (2011). Effects of stimulant medication, incentives, and event rate on reaction time variability in children with ADHD. *Neuropsychopharmacology*, *36*(5), 1060–1072. https://doi.org/10.1038/npp.2010.243

Fang, Y., Forger, D. B., Frank, E., Sen, S., & Goldstein, C. (2021). Day-to-day variability in sleep parameters and depression risk: A prospective cohort study of training physicians. *npj Digital Medicine*, *4*(1), 28. https://doi.org/10.1038/s41746-021-00400-z

Forte, G., Favieri, F., & Casagrande, M. (2019). Heart rate variability and cognitive function: A systematic review. *Frontiers in Neuroscience*, *13*. Retrieved November 24, 2023, from https://www.frontiersin.org/articles/10.3389/fnins.2019.00710

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*(8), 1062–1068. https://doi.org/10.1038/nn.2342

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g?: The relationship between the scholastic assessment test and general cognitive ability [Publisher: SAGE Publications Inc]. *Psychological Science*, *15*(6), 373–378. https://doi.org/10.1111/j.0956-7976.2004.00687.x

Friso-van den Bos, I., van der Ven, S. H. G., Kroesbergen, E. H., & van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review*, *10*, 29–44. https://doi.org/10.1016/j.edurev.2013.05.003

Galeano Weber, E., Dirk, J., & Schmiedek, F. (2018). Variability in the precision of children's spatial working memory. *Journal of Intelligence*, *6*(1), 8. https://doi.org/10.3390/jintelligence6010008

Galeano-Keiner, E. M., Neubauer, A. B., Irmer, A., & Schmiedek, F. (2022). Daily fluctuations in children's working memory accuracy and precision: Variability at multiple time scales and links to daily sleep behavior and fluid intelligence. *Cognitive Development*, *64*, 101260. https://doi.org/10.1016/j.cogdev.2022.101260

Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2004). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age [_eprint: https://onlinelibrary *Applied Cognitive Psychology*, *18*(1), 1–16. https://doi.org/10.1002/acp.934

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, *1*(3). https://doi.org/10.1214/06-BA117A

Grenard, J. L., Ames, S. L., Wiers, R. W., Thush, C., Sussman, S., & Stacy, A. W. (2008). Working memory capacity moderates the predictive effects of drug-related associations on substance use [Place: US Publisher: American Psychological Association]. *Psychology of Addictive Behaviors*, *22*(3), 426–432. https://doi.org/10.1037/0893-164X.22.3.426

Hagemann, D., Ihmels, M., Bast, N., Neubauer, A. B., Schankin, A., & Schubert, A.-L. (2023). Fluid intelligence is (much) more than working memory capacity: An experimental analysis [Number: 4 Publisher: Multidisciplinary Digital Publishing Institute]. *Journal of Intelligence*, *11*(4), 70. https://doi.org/10.3390/jintelligence11040070

Hamilton, H. R., Armeli, S., & Tennen, H. (2023). Too tired to drink? daily associations of sleep duration and fatigue with own and others' alcohol consumption [Publisher: American Psychological Association]. *Psychology of Addictive Behaviors*, *37*(2), 267–274. https://doi.org/10.1037/adb0000882

Hoffman, L., & Walters, R. W. (2022). Catching up on multilevel modeling. *Annual Review of Psychology*, *73*(1), 659–689. https://doi.org/10.1146/annurev-psych-020821-103525

Houben, K., Wiers, R. W., & Jansen, A. (2011). Getting a grip on drinking behavior: Training working memory to reduce alcohol abuse [Publisher: SAGE Publications Inc]. *Psychological Science*, *22*(7), 968–975. https://doi.org/10.1177/0956797611412392

Huang-Pollock, C., Shapiro, Z., Galloway-Long, H., & Weigard, A. (2017). Is poor working memory a transdiagnostic risk factor for psychopathology? *Journal of Abnormal Child Psychology*, *45*(8), 1477–1490. https://doi.org/10.1007/s10802-016-0219-8

Hunt, C. E., Brouillette, R. T., Liu, K., & Klemka, L. (1985). Day-to-day pneumogram variability [Number: 2 Publisher: Nature Publishing Group]. *Pediatric Research*, *19*(2), 174–177. https://doi.org/10.1203/00006450-198502000-00005

Jastrzębski, J., Ciechanowska, I., & Chuderski, A. (2018). The strong link between fluid intelligence and working memory cannot be explained away by strategy use. *Intelligence*, *66*, 44–53. https://doi.org/10.1016/j.intell.2017.11.002

Judd, N., Aristodemou, M., & Kievit, R. (2023, September 1). Cognitive variability is ubiquitous and distinct from mean performance across eleven tasks with over 7 million trials. https://doi.org/10.31234/osf.io/b29rn

Judd, N., & Klingberg, T. (2021). Training spatial cognition enhances mathematical learning in a randomized study of 17,000 children. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01118-4

Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2015, December 29). *Intelligent testing with the WISC-v* [Google-Books-ID: l5dPCwAAQBAJ]. John Wiley & Sons.

Kikuya, M., Ohkubo, T., Metoki, H., Asayama, K., Hara, A., Obara, T., Inoue, R., Hoshi, H., Hashimoto, J., Totsune, K., Satoh, H., & Imai, Y. (2008). Day-by-day variability of blood pressure and heart rate at home as a novel predictor of prognosis: The ohasama study. *Hypertension*, *52*(6), 1045–1050. https://doi.org/10.1161/HYPERTENSIONAHA.107.104620

Kofler, M. J., Rapport, M. D., Sarver, D. E., Raiker, J. S., Orban, S. A., Friedman, L. M., & Kolomeyer, E. G. (2013). Reaction time variability in ADHD: A meta-analytic review of 319 studies. *Clinical Psychology Review*, *33*(6), 795–811. https://doi.org/10.1016/j.cpr.2013.06.001

Kok, A. (2001). On the utility of p3 amplitude as a measure of processing capacity. *Psychophysiology*, *38*(3), 557–577. https://doi.org/10.1017/S0048577201990559

Könen, T., Dirk, J., & Schmiedek, F. (2015). Cognitive benefits of last night's sleep: Daily variations in children's sleep behavior are related to working memory fluctuations. *Journal of Child Psychology and Psychiatry*, *56*(2), 171–182. https://doi.org/10.1111/jcpp.12296

Kouros, C. D., & El-Sheikh, M. (2015). Daily mood and sleep: Reciprocal relations and links with adjustment problems [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.12226]. *Journal of Sleep Research*, *24*(1), 24–31. https://doi.org/10.1111/jsr.12226

Laurence, P. G., & Macedo, E. C. (2023). Cognitive strategies in matrix-reasoning tasks: State of the art. *Psychonomic Bulletin & Review*, *30*(1), 147–159. https://doi.org/10.3758/s13423-022-02160-7

Li, S.-C., von Oertzen, T., & Lindenberger, U. (2006). A neurocomputational model of stochastic resonance and aging. *Neurocomputing*, *69*(13), 1553–1560. https://doi.org/10.1016/j.neucom.2005.06.015

Lin, X., Mermelstein, R. J., & Hedeker, D. (2018). A 3-level bayesian mixed effects location scale model with an application to ecological momentary assessment data. *Statistics in Medicine*, *37*(13), 2108–2119. https://doi.org/10.1002/sim.7627

Lorist, M. M., & Tops, M. (2003). Caffeine, fatigue, and cognition. *Brain and Cognition*, *53*(1), 82–94. https://doi.org/10.1016/S0278-2626(03)00206-9

Ludwig, K., Haindl, A., Laufs, R., & Rauch, W. A. (2016). Self-regulation in preschool children's everyday life: Exploring day-to-day variability and the within- and between-person structure. *Journal of Self-Regulation and Regulation*, *2*, 99–117. https://doi.org/10.11588/josar.2016.2.34357

MacDonald, S. W., Nyberg, L., & Bäckman, L. (2006). Intra-individual variability in behavior: Links to brain structure, neurotransmission and neuronal activity. *Trends in Neurosciences*, *29*(8), 474–480. https://doi.org/10.1016/j.tins.2006.06.011

Nikolin, S., Tan, Y. Y., Schwaab, A., Moffa, A., Loo, C. K., & Martin, D. (2021). An investigation of working memory deficits in depression using the n-back task: A systematic review and meta-analysis. *Journal of Affective Disorders*, *284*, 1–8. https://doi.org/10.1016/j.jad.2021.01.084

Ntekouli, M., Spanakis, G., Waldorp, L., & Roefs, A. (2023, October 11). Model-based clustering of individuals' ecological momentary assessment time-series data for improving forecasting performance. https://doi.org/10.48550/arXiv.2310.07491

Rabbitt, P., Osman, P., Moore, B., & Stollery, B. (2001). There are stable individual differences in performance variability, both from moment to moment and from day to day. *The Quarterly Journal of Experimental Psychology Section A*, *54*(4), 981–1003. https://doi.org/10.1080/713756013

Rast, P., & Ferrer, E. (2018). A mixed-effects location scale model for dyadic interactions. *Multivariate Behavioral Research*, *53*(5), 756–775. https://doi.org/10.1080/00273171.2018.1477577

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? a comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Robinson, P. D., Brown, N. J., Turner, M., Van Asperen, P., Selvadurai, H., & King, G. G. (2014). Increased day-to-day variability of forced oscillatory resistance in poorly controlled or persistent pediatric asthma. *CHEST*, *146*(4), 974–981. https://doi.org/10.1378/chest.14-0288

Roid, G. H., & Pomplun, M. (2012). The stanford-binet intelligence scales, fifth edition. In *Contemporary intellectual assessment: Theories, tests, and issues, 3rd ed* (pp. 249–268). The Guilford Press.

Schmidt, C., Collette, F., Cajochen, C., & Peigneux, P. (2007). A time to think: Circadian rhythms in human cognition. *Cognitive Neuropsychology*, *24*(7), 755–789. https://doi.org/10.1080/02643290701754158

Schmiedek, F., Lövdén, M., von Oertzen, T., & Lindenberger, U. (2020). Within-person structures of daily cognitive performance differ from between-person structures of cognitive abilities. *PeerJ*, *8*, e9290. https://doi.org/10.7717/peerj.9290

Shipstead, Z., Redick, T. S., Hicks, K. L., & Engle, R. W. (2012). The scope and control of attention as separate aspects of working memory [Publisher: Routledge _eprint: https://doi.org/10.1080/09658211.2012.691519]. *Memory*, *20*(6), 608–628. https://doi.org/10.1080/09658211.2012.691519

Solfo, A., & Van Leeuwen, C. (2023). A bayesian classifier for fractal characterization of short behavioral series. *Psychological Methods*. https://doi.org/10.1037/met0000562

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1. https://doi.org/10.1038/s43586-020-00001-2

Varga, S., & Heck, D. H. (2017). Rhythms of the body, rhythms of the brain: Respiration, neural oscillations, and embodied cognition. *Consciousness and Cognition*, *56*, 77–90. https://doi.org/10.1016/j.concog.2017.09.008

Voytek, B., Kramer, M. A., Case, J., Lepage, K. Q., Tempesta, Z. R., Knight, R. T., & Gazzaley, A. (2015). Age-related changes in 1/*f* neural electrophysiological noise. *The Journal of Neuroscience*, *35*(38), 13257–13265. https://doi.org/10.1523/JNEUROSCI.2332-14.2015

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (n.d.). Estimation and interpretation of 1/fa noise in human cognition.