Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-023-02369-0

BRIEF REPORT

Task sequencing does not systematically affect the factor structure of cognitive abilities

Matthew K. Robison[1] · **Xavier Celaya**[2] · **B. Hunter Ball**[1] · **Gene A. Brewer**[2]

Accepted: 13 August 2023

Abstract

Latent variable analyses of cognitive abilities are among the major means by which cognitive psychologists test theories regarding the structure of human cognition. Models are fit to observed variance-covariance structures, and the fit of those models are compared to assess the merits of competing theories. However, an often unconsidered and potentially important methodological issue is the precise sequence in which tasks are delivered to participants. Here we empirically tested whether differences in task sequences systematically affect the observed factor structure. A large sample ($N = 587$) completed a battery of 12 cognitive tasks measuring four constructs: working memory, long-term memory, attention control, and fluid intelligence. Participants were assigned to complete the assessment in one of three sequences: fixed and grouped by construct vs. fixed and interleaved across constructs vs. random by participant. We generated and tested two hypotheses: grouping task sequences by construct (i.e., administering clusters of tasks measuring a cognitive construct consecutively) would (1) systematically increase factor loadings and (2) systematically decrease interfactor correlations. Neither hypothesis was supported. The measurement models were largely invariant across the three conditions, suggesting that latent variable analyses are robust to such subtle methodological differences as task sequencing.

Keywords Individual differences · Working memory · Attention control · Long-term memory · Fluid intelligence

Differential psychologists use the correlational nature of data to make inferences about the nature and structure of psychological traits. In cognitive psychology, differential approaches have a long and important history, dating back to the seminal work of Binet, Cattell, Horn, Thurstone, Thorndike, and Spearman (McGrew, 2009). Latent variable analysis allowed theories to be tested more rigorously and quantitatively. A given model will hypothesize a certain underlying factor structure that is thought to reflect a set of underlying cognitive dimensions or processes. In turn, model fit is assessed by the degree to which a particular factor structure recreates an observed correlational structure, adjusted for parsimony. The findings from such models have nontrivial implications. For example, should we consider working memory a unitary system or distinguish between verbal and nonverbal stores (Kane et al., 2004)? Is intelligence a single construct or better described by multiple types of intelligence (Carroll, 1993)? To what extent does "executive functioning" represent a unitary versus diverse set of functions (Miyake et al., 2000)? What explains the relationship between working memory capacity and fluid intelligence (Draheim et al., 2021; Unsworth et al., 2014; Unsworth & Spillers, 2010)?

Because factor structures are used to test theories, it is worth asking how subtle methodological differences can impact those factor structures. In a typical latent variable design, participants complete a task battery designed to measure a set of cognitive abilities. Multiple tasks of each cognitive ability are administered, and performance on these tasks can be set to load onto factors. The resulting factor structure is then used to test competing predictions made by different theories. Here, we specifically examined one subtle methodological difference: task sequencing. We were motivated by the observation that differential cognitive psychologists rarely randomize the sequencing of tasks in a latent variable analysis. However, this approach seems to violate two critical tenets of experimentation: randomization and counterbalancing.

✉ Matthew K. Robison
matthew.robison@uta.edu

[1] Department of Psychology, The University of Texas at Arlington, Arlington, TX, USA

[2] Department of Psychology, Arizona State University, 501 S. Nedderman Dr., Box 19528, Arlington, TX 76019, USA

Published online: 31 August 2023

To provide an overview of the methods researchers have used in such studies, we examined the Method of a sample of latent variable analyses of cognitive abilities (see Table 1).[1] We categorized task sequencing into three general methods: grouped, interleaved, and random. In a grouped sequence, tasks designed to tap the same cognitive construct (e.g., working memory) are completed consecutively. In an interleaved sequence, the tasks are shuffled such that those measuring the same construct are *not* completed consecutively. In a random sequence, each participant completes the tasks in a different random order. Of the studies reviewed, only two used a random sequence (Colom et al., 2004; Kyllonen & Christal, 1990). Some researchers state that they present the tasks in a fixed sequence to avoid order-related confounds. For example, Miyake et al. (2000) state, "The order of task administration was fixed for all participants (with the constraint that no two tasks that were supposed to tap the same executive function occurred consecutively) to minimize any error due to participant by order interaction" (p. 66). In other instances, the researchers used a grouped sequence, but counterbalanced the order in which the tasks were delivered. For example, Engle et al. (1999) gave measures of working and short-term memory during the first and second days of a 3-day study and gave measures of fluid intelligence on the third day. On Days 1 and 2, they counterbalanced the sequencing of the working and short-term memory tasks across participants. In other cases, consideration was given to the sequencing of tasks, but no differences were observed, and results were reported after collapsing across task orders (Shelton et al., 2009). The main question then is whether these choices lead to different factor structures.

This issue has been considered in other fields, such as survey development and personality assessment. For example, Goodhue and Loiacono (2002) noted that clustering survey items by construct inflates reliability. Other studies have compared randomized item ordering, grouped/clustered orders, and randomized orders (Buchanan et al., 2018; Loiacono & Wilson, 2020; Wilson & Lankton, 2012; Wilson et al., 2017). In a comprehensive examination, Wilson et al. (2021) examined five types of item ordering and clustering in an online survey. They had five conditions which either fixed, interleaved, or randomized the sequence of items in different ways. Wilson et al. found significant differences across orderings in item means, cluster means, reliability, and differences in participant reports of fatigue and frustration. In contrast, Schell and Oswald (2013) assessed 50 Big Five personality items using three item orders—randomized per individual, items grouped by factor, and a fixed order of items interleaved across factors—but did not find any differences in the measurement model or in the internal consistencies of the factors. To our knowledge, no study has systematically evaluated task sequencing in the context of a latent, construct-level analysis of cognitive abilities.[2] That was the central goal of the present study.

If any differences in factor structure should emerge, we hypothesized that a grouped sequence, compared with an interleaved or random sequence, would cause an increase in the magnitude of task loadings onto respective factors (Hypothesis 1) and reduce the magnitude of interfactor correlations (Hypothesis 2). If tasks designed to measure a specific construct are presented consecutively, there will be at least two systematic sources of covariance among them: the latent cognitive ability that causes performance differences on those tasks and their shared temporal variance. In a grouped sequence, any temporal and contextual covariates that might influence one's performance (e.g., fatigue, stress, time of day) will increase the covariance among those tasks. This would be especially true when measures of different constructs are given on different days. Thus, the factor will be a conflation of "true" covariance and temporal/contextual covariance. In turn, the individual factors will have less covariance. Quantitatively, this would manifest in two ways: (1) a systematic *increase* in the magnitude (i.e., absolute value) of factor loadings, and (2) a systematic *decrease* in the magnitude of interfactor correlations. Here, we tested these hypotheses empirically. Specifically, we administered a battery of 12 cognitive tasks with three each selected to measure working memory capacity, attention control, long-term memory, and fluid intelligence. These four constructs have been demonstrated to be distinct yet correlated (Unsworth et al., 2014). We randomly assigned participants to complete the tasks in either a grouped, interleaved, or random sequence. Finally, we tested whether the conditions yielded differences in average performance, latent factor loadings, and interfactor correlations.

Method

We report all variables, how we determined our sample size, and all exclusions, when necessary. All data, materials,[3] and analysis scripts are publicly available on the Open Science Framework (https://osf.io/a79hf/).

[1] This list was not meant to be exhaustive, nor a meta-analysis. Rather, the list was intended to see (1) how often researchers state the order in which their task battery was delivered and (2) if they do state the order, how they sequenced their tasks. In some cases, the task sequence was not explicitly stated in the Method, but we were able to recover that information through personal communication, or because one of the authors of the present study also authored the study.

[2] Although Shelton et al. (2009) report manipulating task order across participants, they do not report their analyses on the differences or lack thereof.

[3] With the exception of the Raven matrices, which are proprietary.

Table 1 Task sequences in a sample of latent-variable analysis of cognitive abilities

Study	Constructs	Defined	Task sequence
Ackerman et al. (2002)	WMC, PS, gF	No	WMC grouped
Brewer and Unsworth (2012)	WMC, AC, LTM, gF	Yes	WMC grouped
Chuderski and Neçka (2012)	WMC, AC, STM, gF	Yes	Grouped
Colom et al. (2004)	WMC, STM, PS, gF, gC	Yes	Random
Colom et al. (2005)	WMC, STM, gF	Yes	Interleaved
Conway et al. (2002)	WMC, STM, PS, gF	Yes	gF grouped
Cowan et al. (2005)	WMC, STM, gF	Yes	Grouped
Engle et al. (1999)	WMC, STM, gF	Yes	Grouped
Frith et al. (2021)	AC, gF, Creativity	No	Grouped
Kane et al. (2004)	WMC, STM, gF	Yes	Grouped
Kane et al. (2016)	WMC, AC	Yes	Interleaved
Kyllonen and Christal (1990)	WMC, gF, gC, PS	Yes	Random
Miyake et al. (2000)	EF	Yes	Interleaved
Miyake et al. (2001)	WMC, STM, gS, EF	Yes	Interleaved
Oberauer et al. (2000)	WMC, PS	No	Interleaved
Redick et al. (2016)	WMC, AC, gF, MT	Yes	Interleaved
Rey-Mermet et al. (2019)	WMC, AC, gF	No	Interleaved
Robison and Brewer (2020)	WMC, STM	Yes	Interleaved
Robison and Brewer (2022)	WMC, AC, gF	Yes	Interleaved
Robison and Unsworth (2018)	WMC, AC	No	Grouped
Robison et al. (2017)	WMC, AC, gF	No	Grouped
Robison et al. (2020)	WMC, AC	Yes	Grouped
Shelton et al. (2009, 2010)	WMC, AC, STM, LTM, gF	Yes	Interleaved
Shipstead et al. (2014)	WMC, AC, STM, LTM, gF	Yes	Interleaved
Shipstead et al. (2015)	WMC, AC, STM, gF	Yes	Interleaved
Tsukahara et al. (2020)	WMC, AC, SD	Yes	Interleaved
Unsworth (2010)	WMC, LTM, gC, gF	Yes	Grouped
Unsworth, Brewer, et al. (2009a)	WMC, AC, gF	Yes	Grouped
Unsworth, Spillers, et al. (2009c)	WMC, LTM, gF	Yes	Grouped
Unsworth, Redick, et al. (2009b)	WMC, gF	Yes	Grouped
Unsworth and Spillers (2010)	WMC, AC, LTM, gF	Yes	WMC/AC grouped
Unsworth et al. (2010)	WMC, AC, gF, Fluency	Yes	WMC/gF grouped
Unsworth et al. (2012)	WMC, AC, LTM, PM	Yes	WMC grouped
Unsworth et al. (2014)	WMC, AC, STM, LTM, gF	Yes	Interleaved
Unsworth and McMillan (2014)	WMC, AC, gF	Yes	Grouped
Unsworth and McMillan (2017)	WMC, AC, gF	Yes	Grouped
Unsworth et al. (2019)	WMC, AC	Yes	Grouped
Unsworth et al. (2021)	WMC, AC	Yes	Grouped

WMC = working memory capacity; AC = attention control; STM = short-term memory; LTM = long-term memory; gF = fluid intelligence; gC = crystallized intelligence; PS = processing speed; EF = executive functioning; MT = multitasking; SD = sensory discrimination.

Participants and procedure

A priori**,** we targeted a sample size of 600 participants (200 per condition) based on simulations from Kretzschmar and Gignac (2019). We used the end of an academic semester as our stopping rule for data collection, and we finished just short of our target sample size with 598 participants. After exclusions (see below), the final sample analyzed included 587 participants (172 in grouped condition, 217 in interleaved condition, and 198 in random condition; 44% participants identified as women, 54% as men, 1% as nonbinary or other gender, and one participant did not to report gender; M_{age} = 19.01 years, SD_{age} = 1.49, range: 17–37; 12% of participants identified as Asian, 6% as Black or African American, 1% as Native Hawaiian or Pacific Islander, 21% as Hispanic or Latino, 55% as White, and 4% as other race/

ethnicity). All participants were undergraduate students at Arizona State University who completed the study in exchange for partial course credit. Participants completed the study in groups of four to eight. Prior to beginning the study, participants read and signed an informed consent document. The experimental protocol was approved by the Institutional Review Board at Arizona State University.

We created an R script that first randomly assigned a subject ID number to a condition. For the random condition, the R script took the 12 task labels and shuffled them. For all conditions, the script printed the task sequence onto checklists (see Fig. 1 for task orders by condition). The research assistant then used this printed checklist to administer the task sequence for each participant in precisely the order listed on the sheet. All sessions took approximately 2 hours to complete.

Tasks

Working memory capacity

Operation span (Unsworth et al., 2005) Participants were required to remember and recall lists of letters while solving math problems as a secondary processing task. On each list, a single letter appeared for 1 s, followed by a math problem (e.g., $(2 \times 5) + 3 = ?$). The participant clicked the mouse to indicate that they had solved the problem. Then, they were shown a solution, and they clicked boxes labeled "true" or "false" to indicate whether the solution solved the problem. The process repeated for a list of three to seven items. Each list length was presented twice. At the end of a list, participants were presented with a grid of the possible letters, and their task was to report the letters in correct forward serial order. The dependent variable was the total number of letters reported in the correct serial position (maximum score = 50). [4]

Symmetry span (Unsworth, Redick, et al., 2009b) Participants were required to remember sequences of spatial locations while making symmetry judgments as a secondary processing task. On each list, a single location within a 4×4 black-and-white grid flashed red for 500 ms. Then, participants were presented with a black-and-white pattern, and their task was to determine whether the pattern was symmetrical about its *y*-axis. When they had made their judgment, they clicked the mouse. Then clicked one of two boxes labeled "true" or "false." This process repeated for two to five items. Each list length was presented twice. At the end of each list, the participants were presented with an empty 4×4 grid and asked to click the locations that appeared on the list in forward serial order. The dependent variable was the total number of locations reported in the correct serial position (maximum score = 28).

Reading span (Unsworth, Redick, et al., 2009b) Participants were required to remember and recall lists of letters while making judgments about the sensibleness of sentences as a secondary processing task. On each list, a single letter appeared for 1 s, followed by a sentence (e.g., Jack went up the hill to fetch a trash of water). The participant clicked the mouse to indicate that they had made the sentence judgment. Then, they were shown boxes labeled "true" and "false" and clicked a box to indicate whether the sentence made sense or not. The process repeated for a list length of three to seven items. Each list length was presented twice. At the end of a list, participants were presented with a grid of the possible letters, and their task was to report the letters in correct forward serial order by clicking a box next to each letter. The dependent variable was the total number of letters reported in the correct serial position (maximum score = 50).

Attention control

Antisaccade (Hutchison, 2007; Kane et al., 2001) On each trial, a central fixation stimulus (***) appeared for either 1 or 2 s. Then, a flashing white cue (=) appeared on either the right or left side of the screen for 300 ms. A target letter (*O* or *Q*) then flashed for 100 ms on the opposite side of the screen, followed by a backward mask (#) until the participant made their response. Participant made their responses with the *O* and *Q* keys of the keyboard. The next trial started after 1-s blank intertrial interval. Participants first received 8 trials of slow-paced practice, in which the target appeared for 500 ms, then 16 trials of fast-paced practice with a 100-ms target duration, and finally 72 experimental trials. The dependent variable was proportion correct on the experimental trials.

PVT (Dinges & Powell, 1985; Wilkinson & Houghton, 1982) On each trial, a millisecond counter appeared at the center of the screen (00.000). Then, after a random time interval ranging from 2 to 8 s, the timer began counting like a stopwatch. The participant's task was to press the spacebar as quickly as possible when the timer started. There were five practice trials followed by 75 experimental trials. Reaction times across experimental trials were then sorted from fastest to slowest for each participant and binned into quintiles. The dependent variable was the average of each participants slowest quintile (Unsworth & Spillers, 2010).

SART (Robertson et al., 1997) On each trial, a single digit (1–9) appeared at the center of the screen for 300 ms, followed by a 900-ms blank intertrial interval. The participant's

[4] Although it is common to use exclude participants with less than 85% accuracy on complex span tasks, we opted not to use this threshold, as a recent study indicated it tends to disproportionately eliminate low working memory individuals (Richmond et al., 2021).

Fig. 1 Task sequences by condition. The sequence in the grouped and interleaved conditions was fixed for all participants, whereas the sequence in the random condition was different for every participant. "*" represents an example task sequence

task was to press the spacebar upon seeing any digit except 3. Participants were instructed to withhold their response upon seeing the digit 3. There were 450 trials, 11% of which were "no-go" trials. The dependent variable was the standard deviation of reaction times on "go" trials.

Long-term memory

Delayed free recall Participants were presented with 10 lists of 10 words each. Each word appeared for 1 s each, separated by a 500-ms blank interval. At the end of the list, participants were given math problems to solve for 16 s. Then, they were given 45 s to recall as many words from the previous list as possible. The dependent variable was the average proportion of words recalled per list.

Picture source-recognition Participants were presented with images and asked to remember their spatial location. During the study phase, 30 images were presented for 3 s each, separated by a 500-ms blank interval. Images appeared in one of four screen quadrants (top left, top right, bottom left, bottom right). During the test phase, participants were presented with 60 images (the 30 old images and 30 new images). Participants were asked to report whether the item was new or old. And, if the image was old, the participant needed to report the quadrant in which it was presented. Participants used the number pad on the keyboard to make their responses (1 = old image from bottom left, 3 = old image from bottom right, 7 = old image from top left, 9 = old image from top right, 5 = new image). Due to a programming error, accuracy was only correctly recorded for the new images. Therefore, the dependent variable was accuracy on the new trials (i.e., correct rejections).

Cued recall Participants were presented with five lists of 10 unrelated cue–target word pairs (e.g., horse–gift). Each pair was presented for 2 s followed by a 1-s blank interval. Then, during the test phase, participants were presented with the cue word (e.g., horse–???) and asked to recall the target word it was paired cue during study. Participants were given a maximum of 5 s to recall each target before the next cue was presented. Cues were presented in a different random order than during study. The dependent variable was the average proportion of correctly recalled target words across the five lists.

Fluid intelligence

Raven advanced progressive matrices (Raven & Court, 1962) On each trial, participants were presented with a 3×3 grid of patterns. The bottom-right piece to each pattern was missing. Below the grid, eight possible solutions were provided. The participants' task was to select the solution that best completed an implicit pattern in the grid. Participants completed the 18 odd-numbers problems. Participants were given 10 min to solve as many problems as possible. The dependent variable was the total number of correctly solved problems.

Number series (Thurstone, 1938) On each trial, participants were presented with a sequence of numbers. The task was to select from a set of 5 possible options the number that best continued an implicit pattern in the sequence. Participants were given 4.5 minutes to solve as many of 15 items as possible. The dependent variable was the total number of correctly solved problems.

Letter sets (Ekstrom & Harman, 1976) On each trial, participants were given five sets of four letters. The task was to select the letter set that did not follow a rule present in the other items. Participants were given 5 minute to solve as many of 20 possible problems as possible. The dependent variable was the total number of correctly solved problems.

Table 2 Zero-order correlations among measures in full sample

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. Operation span	–										
2. Symmetry span	**0.34**	–									
3. Reading span	**0.46**	**0.28**	–								
4. Antisaccade	**0.19**	**0.36**	**0.25**	–							
5. Psychomotor vigilance	**−0.19**	**−0.14**	**−0.37**	**−0.32**	–						
6. SART	**−0.10**	**−0.11**	**−0.17**	**−0.29**	**0.27**	–					
7. Delayed free recall	**0.26**	**0.20**	**0.29**	**0.21**	**−0.20**	**−0.22**	–				
8. Picture source-recognition	0.03	**0.12**	**0.15**	**0.13**	**−0.10**	**−0.18**	**0.16**	–			
9. Cued recall	**0.17**	**0.10**	**0.25**	**0.14**	**−0.16**	**−0.18**	**0.45**	**0.26**	–		
10. Raven	**0.23**	**0.27**	**0.29**	**0.34**	**−0.30**	**−0.22**	**0.36**	**0.21**	**0.32**	–	
11. Letter sets	**0.20**	**0.25**	**0.25**	**0.31**	**−0.18**	**−0.21**	**0.28**	**0.12**	**0.25**	**0.33**	–
12. Number series	**0.27**	**0.31**	**0.28**	**0.35**	**−0.24**	**−0.16**	**0.24**	0.06	**0.18**	**0.39**	**0.34**

SART = Sustained Attention to Response Task. Boldface correlations are significant at $p < .05$.

Data analysis

The data were aggregated in R with the *tidyverse* (Wickham et al., 2019), *papaja* (Aust, 2023; Aust & Barth, 2018), and *data.table* (Dowle & Srinivasan, 2021) set of packages, plotted using *ggplot* (Wickham, 2016), *cowplot* (Wilke et al., 2019), and *ggrain* (Allen et al., 2021), and analyzed with the *lavaan* (Rosseel, 2012) and *rstatix* (Kassambara, 2020) packages. The analysis script is available on the Open Science Framework (https://osf.io/qe2kw/). To account for missing data in the latent-variable model fitting, we used maximum-likelihood estimation, which allows all available pairwise relations to inform the variance-covariance matrix to which the model is fit.

Exclusions

We used an outlier detection threshold of 2.5 standard deviations outside each variable's mean to remove extreme values and ensure multivariate normality. Any value falling outside this range was set to missing for the analysis. Proportions of missing/excluded data for each variable are listed in Table 6.[5]

Results

Zero-order correlations among the measures for the full sample are listed in Table 2, and then listed by condition in Tables 3, 4, and 5. Descriptive statistics for the full sample are listed in Table 6. The distributions of task performance by condition are shown in Fig. 2. In our first set of comparisons, we tested for differences in the zero-order correlations between conditions. The comparisons were conducted using Fisher's *r* to *z* transformation, then performing Fisher's (1925) test for a significant difference between correlations measured in independent samples. Six out of 198 comparisons reached significance at $p < .05$. However, about 10 (198×0.05) differences would be expected by chance. No correlations reached the critical threshold of $p < .001$ (see Supplemental Materials for tables of *p* values for each comparison). [6,7] Therefore, the comparisons of zero-order correlations did not suggest a systematic strengthening or weakening of the correlations, either within or across constructs, based on task sequencing Table 7.

Measurement invariance

Our next set of analyses examined whether task sequencing had any impact on the factor structure. First, we specified a confirmatory factor analysis with the operation span, symmetry span, and reading span tasks loading onto a *Working Memory* factor, antisaccade, psychomotor vigilance, and SART loading onto an *Attention Control* factor, delayed free recall, picture source-recognition, and cued recall loading onto a *Long-Term Memory* factor, and Raven, number series, and letter sets loading onto a *Fluid Intelligence* factor (see Fig. 3 for a visualization of the factor structure). This model fit the data well, $\chi^2(48) = 135.68$, CFI = 0.94, TLI = 0.91, RMSEA = 0.055 90%

[5] We also performed a mini multiverse analysis on the data with various types of outlier thresholds, and each yielded qualitatively similar results. Therefore, we felt a standardized ±2.5 *SD* threshold for detecting and removing outlying values was the simplest and most straightforward correction to the data.

[6] A Bonferroni correction to the α would have created a threshold of 0.0008 (0.05/66).

[7] We used G*Power 3.1 to conduct a post hoc power analysis (Faul et al., 2009). With the achieved sample sizes in each condition, we had 80% power to detect a difference in the correlations of about |0.25|—a medium-sized effect.

Table 3 Zero-order correlations among measures in grouped condition

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. Operation span	–										
2. Symmetry span	**0.47**	–									
3. Reading span	**0.40**	**0.29**	–								
4. Antisaccade	**0.26**	**0.43**	**0.29**	–							
5. Psychomotor vigilance	**−0.17**	−0.11	**−0.30**	**−0.45**	–						
6. SART	−0.06	−0.14	**−0.19**	**−0.31**	**0.29**	–					
7. Delayed free recall	**0.22**	**0.16**	**0.26**	**0.22**	**−0.27**	**−0.28**	–				
8. Picture source-recognition	0.09	**0.16**	**0.28**	**0.24**	−0.13	**−0.29**	**0.31**	–			
9. Cued recall	**0.18**	**0.24**	**0.30**	**0.17**	**−0.28**	**−0.28**	**0.52**	**0.37**	–		
10. Raven	**0.19**	**0.24**	**0.20**	**0.34**	**−0.27**	**−0.25**	**0.34**	**0.31**	**0.39**	–	
11. Letter sets	0.15	**0.28**	**0.30**	**0.32**	**−0.31**	**−0.24**	**0.26**	**0.24**	**0.30**	**0.37**	–
12. Number series	**0.21**	**0.24**	**0.18**	**0.17**	**−0.22**	**−0.21**	**0.27**	0.12	**0.26**	**0.30**	**0.35**

SART = Sustained Attention to Response Task. Boldface correlations are significant at $p < .05$.

CI [0.045, 0.067], SRMR = 0.04. Next, we reestimated the model, allowing the factor loadings and interfactor correlations to be estimated for each condition individually. Table 8 lists factor loadings and interfactor correlations by condition. The model fit the data well in in the *grouped* condition, $\chi^2(48) = 63.18$, CFI = 0.96, TLI = 0.95, RMSEA = 0.042 90% CI [0.00, 0.068], SRMR = 0.05 and the *interleaved* condition, $\chi^2(48) = 64.67$, CFI = 0.96, TLI = 0.95, RMSEA = 0.040 90% CI [0.00, 0.063], SRMR = 0.05, but fit slightly worse in the *random* condition, $\chi^2(48) = 84.29$, CFI = 0.91, TLI = 0.88, RMSEA = 0.06 90% CI [0.04, 0.08], SRMR = 0.05.

Factor loadings

Next, we added an equality constraint to the factor loadings. Our first test compared the grouped condition to the combination of the interleaved and random conditions. We added a specification to the model that constrained all factor loadings to be equal across those two groups. Doing so produced a significantly worse-fitting model, according to the comparison of χ^2 fit indices, $\Delta\chi^2(8) = 18.06$, $p = .02$. However, a Bayes factor comparison heavily favored a simpler model in which the factor loadings were fixed across groups (BF > 100,000). Thus, there was evidence against a difference in factor loadings. Next, we added the same equality constraint on the loadings across all three conditions. This model also fit significantly worse than the freely estimated model, based on a χ^2 comparison, $\Delta\chi^2(16) = 35.54$, $p = .001$. However, a Bayes factor comparison heavily favored a simpler model in which all factor loadings were fixed, BF > 100,000. To delve into any specific difference(s), we iteratively compared models by fixing one factor loading at a time to be equal across the grouped and interleaved/random conditions, then between each condition individually.

Table 4 Zero-order correlations among measures in interleaved condition

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. Operation span	–										
2. Symmetry span	**0.38**	–									
3. Reading span	**0.46**	**0.31**	–								
4. Antisaccade	**0.16**	**0.28**	**0.24**	–							
5. Psychomotor vigilance	**−0.21**	−0.13	**−0.41**	**−0.24**	–						
6. SART	−0.08	−0.11	**−0.18**	**−0.38**	**0.31**	–					
7. Delayed free recall	**0.33**	**0.21**	**0.34**	**0.24**	**−0.19**	**−0.19**	–				
8. Picture source-recognition	0.11	0.10	**0.18**	0.05	−0.06	**−0.17**	0.13	–			
9. Cued recall	**0.20**	0.01	**0.25**	**0.17**	**−0.15**	−0.14	**0.43**	**0.24**	–		
10. Raven	**0.27**	**0.23**	**0.36**	**0.36**	**−0.31**	**−0.24**	**0.41**	**0.15**	**0.31**	–	
11. Letter sets	**0.23**	**0.28**	**0.30**	**0.33**	**−0.15**	**−0.24**	**0.37**	0.08	**0.20**	**0.32**	–
12. Number series	**0.28**	**0.35**	**0.38**	**0.37**	**−0.23**	**−0.17**	**0.28**	0.05	**0.15**	**0.46**	**0.37**

SART = Sustained Attention to Response Task. Boldface correlations are significant at $p < .05$.

Table 5 Zero-order correlations among measures in random condition

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. Operation span	–										
2. Symmetry span	**0.18**	–									
3. Reading span	**0.51**	**0.25**	–								
4. Antisaccade	**0.17**	**0.40**	**0.23**	–							
5. Psychomotor vigilance	**−0.17**	**−0.18**	**−0.40**	**−0.31**	–						
6. SART	−0.14	−0.08	**−0.15**	**−0.17**	**0.18**	–					
7. Delayed free recall	**0.22**	**0.22**	**0.27**	**0.16**	**−0.16**	**−0.20**	–				
8. Picture source-recognition	−0.12	0.11	0.00	0.12	−0.11	−0.10	0.07	–			
9. Cued recall	0.13	−0.04	**0.19**	0.08	−0.06	**−0.16**	**0.41**	**0.20**	–		
10. Raven	**0.22**	**0.33**	**0.28**	**0.32**	**−0.33**	**−0.18**	**0.32**	**0.19**	**0.28**	–	
11. Letter sets	**0.20**	**0.21**	0.13	**0.30**	−0.09	−0.12	**0.17**	0.06	**0.31**	**0.30**	–
12. Number series	**0.30**	**0.32**	**0.23**	**0.48**	**−0.27**	−0.13	**0.17**	0.01	**0.15**	**0.39**	**0.29**

SART = Sustained Attention to Response Task. Boldface correlations are significant at $p < .05$.

There is only one degree of freedom in these model comparisons, which allowed us to test whether a specific factor loading differed across conditions. Because there were 48 comparisons, we adjusted our α level for these tests to 0.001 (0.05/48). At this threshold, only one comparison produced a significantly worse model when the parameter was fixed: the loading for cued recall onto the Long-Term Memory factor was higher in the grouped condition than in the other two conditions (see Table 8). We also compared the loadings via Bayes factors. Again, there was only substantial evidence for a difference in the picture source-recognition loading between the grouped and ungrouped conditions. In almost all other cases, there was substantial evidence ($BF_{10} > 3$) against a difference (see Table S2). Thus overall, we found evidence against Hypothesis 1—that administering the tasks in a fixed sequence that groups them by construct would inflate their factor loadings.

Interfactor correlations

Our next test of measurement invariance examined whether the conditions systematically decreased the magnitude of the interfactor correlations. We did not have a hypothesis for a difference between the interleaved and random conditions. Therefore, our first test compared the grouped condition to the interleaved and random conditions combined. We specified an equality constraint on the latent covariances. Doing so did not significantly worsen fit, $\Delta\chi^2(6) = 11.30$, $p = .08$. A Bayes factor comparison heavily favored a simpler model in which all latent covariances were set to be equal

Table 6 Descriptive statistics in full sample

Measure	*N*	Mean	*SD*	Skew	Kurtosis	Reliability	% excluded
Operation span	558	39.17	7.91	−0.75	−0.05	0.72	6
Symmetry span	560	21.07	4.86	−0.68	0.08	0.63	6
Reading span	569	35.98	9.12	−0.70	0.00	0.75	4
Antisaccade	576	0.76	0.14	−0.47	−0.83	0.90	3
Psychomotor vigilance	520	510.87	116.75	1.30	1.85	0.87	12
SART	560	123.33	64.12	1.01	0.37	0.96	6
Delayed free recall	565	0.36	0.16	−0.18	−0.01	0.92	5
Picture source-recognition	556	0.60	0.15	−1.37	1.32	0.94	6
Cued recall	562	0.30	0.18	0.65	−0.44	0.88	5
Raven	581	8.30	3.49	−0.19	−0.61	0.84	2
Letter sets	571	8.44	2.75	0.03	−0.33	0.73	4
Number series	558	8.39	2.86	−0.05	−0.67	0.76	6

SD = standard deviation; SART = Sustained Attention to Response Task. For the complex span tasks, reliability was estimated using a Cronbach's α on accuracy by set size. For all other tasks, reliability was estimated by correlating odd and even trials and applying the Spearman–Brown split-half correction to the correlation.

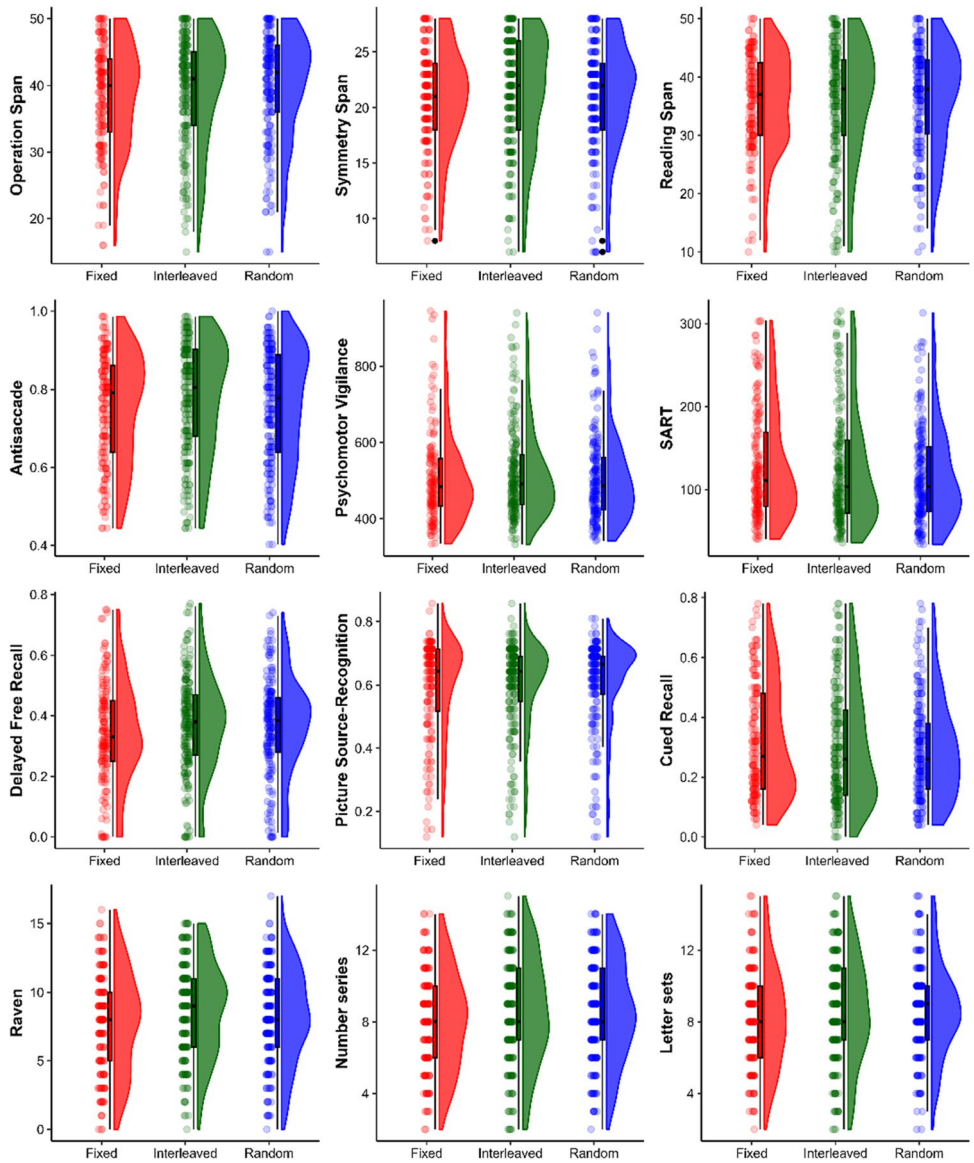

Fig. 2 Task performance by condition

(BF > 100,000). For completeness, we then performed the full condition-wide comparison, fixing the interfactor correlations to be equal across all three conditions. Again, this did not yield a significant worsening of model fit, $\Delta\chi^2(12) = 16.73$, $p = 0.16$, and the Bayes factor comparison heavily favored a simpler model in which all latent correlations were fixed across the three conditions (BF > 100,000). Therefore, we also found evidence against Hypothesis 2—that sequencing the tasks by construct would systematically decrease the interfactor correlations. Although the interleaving and randomization did appear to increase some latent correlations (see Table 8), there was not a systematic decrease in latent correlations, as hypothesized. Figure 3 shows scatterplots of interfactor correlations by condition.

As a final test of our hypothesis, we used the data from the interleaved condition to test whether accounting for temporal proximity would improve model fit. As a reminder, the task sequence in the interleaved condition was the same for all participants. The task order was operation span, antisaccade, delayed free recall, Raven, symmetry span, psychomotor vigilance, picture source-recognition, letter sets, reading span, SART, cued recall, number series. First, we estimated the confirmatory factor analysis. Then, we allowed the residual variances from neighboring tasks (e.g., operation span and antisaccade, antisaccade and delayed free recall) to correlate. Doing so did not improve model fit, $\Delta\chi^2(11) = 9.36$, $p = .59$, and a Bayes factor comparison heavily favored a simpler

Table 7 Factor loadings and interfactor correlations for each condition

Factor loadings				
			Condition	
Factor	Measure	Grouped	Interleaved	Random
Working memory	Operation span	0.64 (0.07)	0.64 (0.06)	0.64 (0.07)
	Symmetry span	0.67 (0.07)	0.52 (0.07)	0.47 (0.08)
	Reading span	0.62 (0.08)	0.75 (0.05)	0.76 (0.07)
Attention control	Antisaccade	0.67 (0.07)	0.66 (0.07)	0.64 (0.07)
	Psychomotor vigilance	−0.68 (0.07)	−0.49 (0.08)	−0.54 (0.07)
	SART	−0.53 (0.07)	−0.54 (0.07)	−0.30 (0.08)
Long-term memory	Delayed free recall	0.70 (0.06)	0.77 (0.07)	0.65 (0.09)
	Picture source-recognition	0.52 (0.07)	0.24 (0.09)	0.21 (0.09)
	Cued recall	0.76 (0.06)	0.56 (0.07)	0.65 (0.09)
Fluid intelligence	Raven	0.66 (0.06)	0.67 (0.05)	0.64 (0.06)
	Number series	0.50 (0.08)	0.63 (0.06)	0.61 (0.06)
	Letter sets	0.62 (0.07)	0.56 (0.06)	0.47 (0.07)
Interfactor correlations				
			Condition	
Factor	Correlate	Grouped	Interleaved	Random
Working memory	Attention control	0.62 (0.10)	0.58 (0.10)	0.73 (0.11)
Working memory	Long-term memory	0.54 (0.09)	0.63 (0.08)	0.49 (0.10)
Working memory	Fluid intelligence	0.65 (0.10)	0.77 (0.07)	0.68 (0.10)
Attention control	Long-term memory	0.62 (0.09)	0.49 (0.10)	0.42 (0.12)
Attention control	Fluid intelligence	0.79 (0.09)	0.80 (0.08)	0.95 (0.09)
Long-term memory	Fluid intelligence	0.78 (0.08)	0.72 (0.08)	0.61 (0.10)

SART = Sustained Attention to Response Task. Values represent standardized loadings and interfactor correlations. Standard errors are listed in parentheses.

model without adding residual covariances between adjacent tasks (BF > 100,000). Therefore, we again did not find evidence that measures delivered near each other in time systematically shared variance Fig. 4.

Mean differences

Although we did not have any specific hypotheses regarding mean differences across conditions, we submitted the data to one-way analyses of variance (ANOVAs) with a between-subjects factor for condition. Because we estimated 12 ANOVAs, we adjusted our α level to correct for multiple comparisons (0.05/12 = 0.004). At this threshold, no ANOVAs indicated significant differences in mean performance (see Table 9).

Next, we used factor analysis to compare construct-level means. The factor scores were saved for each participant. Factor scores were normally distributed with |skew| values < 1 and |kurtosis| values < 1.50. The scores were submitted to one-way ANOVAs with a between-subjects factor of condition. There were no significant differences in average factor scores across conditions: Working Memory: $F(2, 591) = 0.58$, $p = 0.56$, $\eta^2 = 0.002$, $BF_{01} = 333.28$; Attention Control: $F(2, 591) = 1.95$, $p = 0.14$, $\eta^2 = 0.007$, $BF_{01} = 84.55$; Long-Term Memory: $F(2, 591) = 0.58$, $p = 0.56$, $\eta^2 = 0.002$, $BF_{01} = 330.45$; Fluid Intelligence: $F(2, 584) = 2.75$, $p = 0.07$, $\eta^2 = 0.009$, $BF_{01} = 38.11$; see Fig. 5.

Discussion

The present study was motivated by the observation that differential cognitive psychologists rarely randomize the sequencing of tasks in a latent variable analysis, which violates the principles of randomization and counterbalance in experimental psychology. Indeed, we often encounter a critique that our latent-variable designs are confounded by delivering tasks in fixed orders. This has been a concern of considerable deliberation in other fields, such as survey design (Buchanan et al., 2018; Loiacono & Wilson, 2020; Schell & Oswald, 2013; Wilson & Lankton, 2012; Wilson et al., 2017, 2021). Our goal in the present study was to test whether task sequencing systematically affects the latent factor structure of cognitive abilities. This is a nontrivial issue, as factor specification and correlations among factors are used to test theories regarding the structure of cognition.

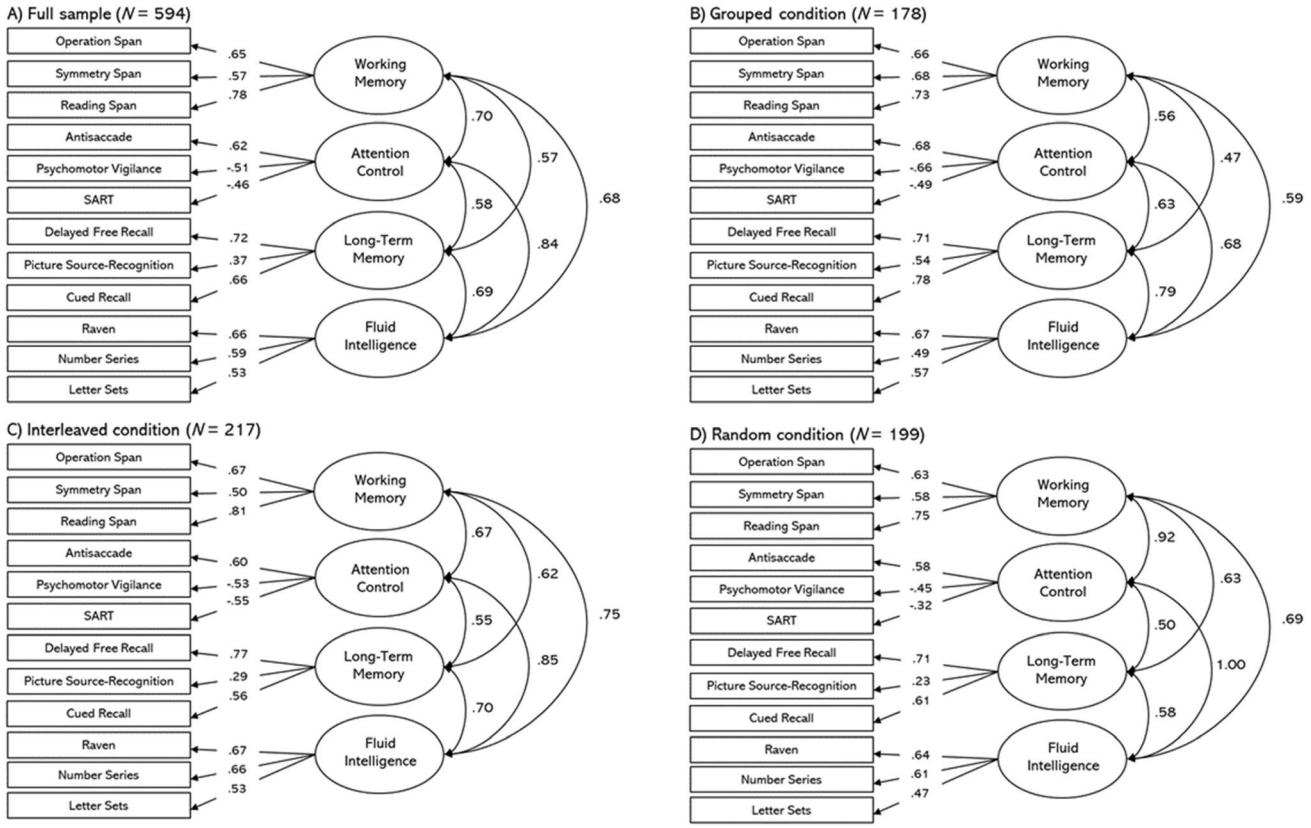

Fig. 3 Factor loadings and interfactor correlations for **A)** full sample, **B)** grouped condition, **C)** interleaved condition, and **D)** random condition. All parameters are standardized, and all were significant at $p < .05$

We tested two hypotheses for how grouping cognitive tasks would affect the factor structure: compared with interleaved and randomized task sequences, grouped task sequences would (1) inflate factor loadings and (2) constrict interfactor correlations. Both effects were hypothesized to occur because of shared temporal variance being conflated

Table 8 Results of model comparisons fixing individual factor loadings

	Comparison			
	Grouped vs. Ungrouped*	Grouped vs. Interleaved	Grouped vs. Random	Interleaved vs. Random
Fixed loading	$\Delta\chi^2(1)$	$\Delta\chi^2(1)$	$\Delta\chi^2(1)$	$\Delta\chi^2(1)$
Operation span	0.05	0.08	0.02	0.02
Symmetry span	1.00	0.39	1.26	0.31
Reading span	4.70	4.79	2.44	0.30
Antisaccade	0.01	0.00	0.00	0.00
Psychomotor vigilance	3.00	2.79	2.22	0.04
SART	1.42	0.14	**5.36**	**7.02**
Delayed free recall	0.05	0.03	0.49	0.77
Picture source-recognition	**9.28**	**6.45**	**7.43**	0.06
Cued recall	3.92	2.70	2.30	0.00
Raven	0.30	0.22	0.40	0.04
Number series	2.00	2.29	1.02	0.32
Letter sets	1.44	0.22	2.83	1.75

Boldface χ^2 values are significant at $p < .05$. *The random and interleaved conditions were combined to represent the "ungrouped" task sequence.

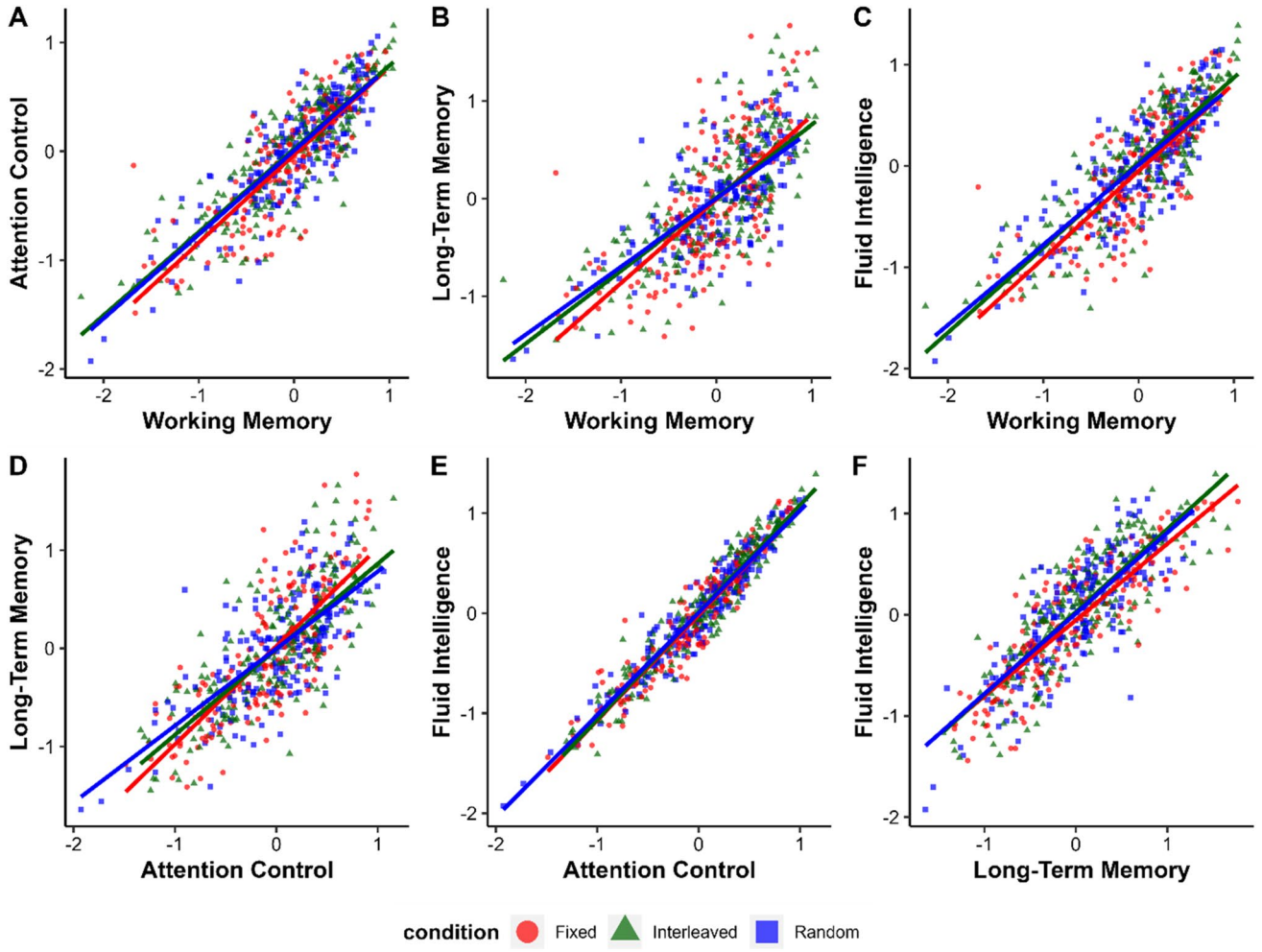

Fig. 4 Scatterplots of interfactor correlations. Points and lines of best fit are plotted separately for each condition. (Color figure online)

with "true" shared variance – that due to a common underlying cognitive ability. Overall, the data did not show any systematic effects of task sequencing on the factor structure. Although the χ^2 comparisons yielded significant worsening of model fit by fixing factor loadings, a Bayes factor comparison heavily favored a more parsimonious model in which all factor loadings were equal across all conditions. This was evidence against Hypothesis 1—that organizing the task sequence by construct would increase factor loadings. Further, there were no differences across the conditions in the magnitude of the interfactor correlations. Therefore, we cannot conclude that task sequencing is a significant moderator of the coherence of putative measures of a cognitive ability within a factor, nor that it moderates the strength of correlations among latent factors.

There are a few limitations worth mentioning. First, we gave the 12 tasks on the same day during a single 2-hr session. Therefore, the temporal grouping was still quite narrow. It is not uncommon for large batteries to be completed across multiple days. In that case, the shared temporal context for same-day tasks versus different-day tasks might be much stronger. Therefore, the present results may not generalize

Table 9 ANOVAs on dependent variables by condition

Measure	*F*	*df*	*p*	η^2
Operation span	0.22	(2, 565)	.81	0.001
Symmetry span	0.06	(2, 572)	.94	<0.001
Reading span	0.38	(2, 578)	.68	0.001
Antisaccade	1.59	(2, 539)	.21	0.006
Psychomotor vigilance	0.06	(2, 530)	.94	<0.001
SART	2.94	(2, 568)	.05	0.010
Delayed free recall	0.41	(2, 541)	.66	0.002
Picture source-recognition	0.91	(2, 540)	.40	0.003
Cued recall	1.40	(2, 562)	.25	0.005
Raven	1.06	(2, 541)	.35	0.004
Number series	1.22	(2, 545)	.30	0.004
Letter sets	0.67	(2, 565)	.51	0.002

SART = Sustained Attention to Response Task.

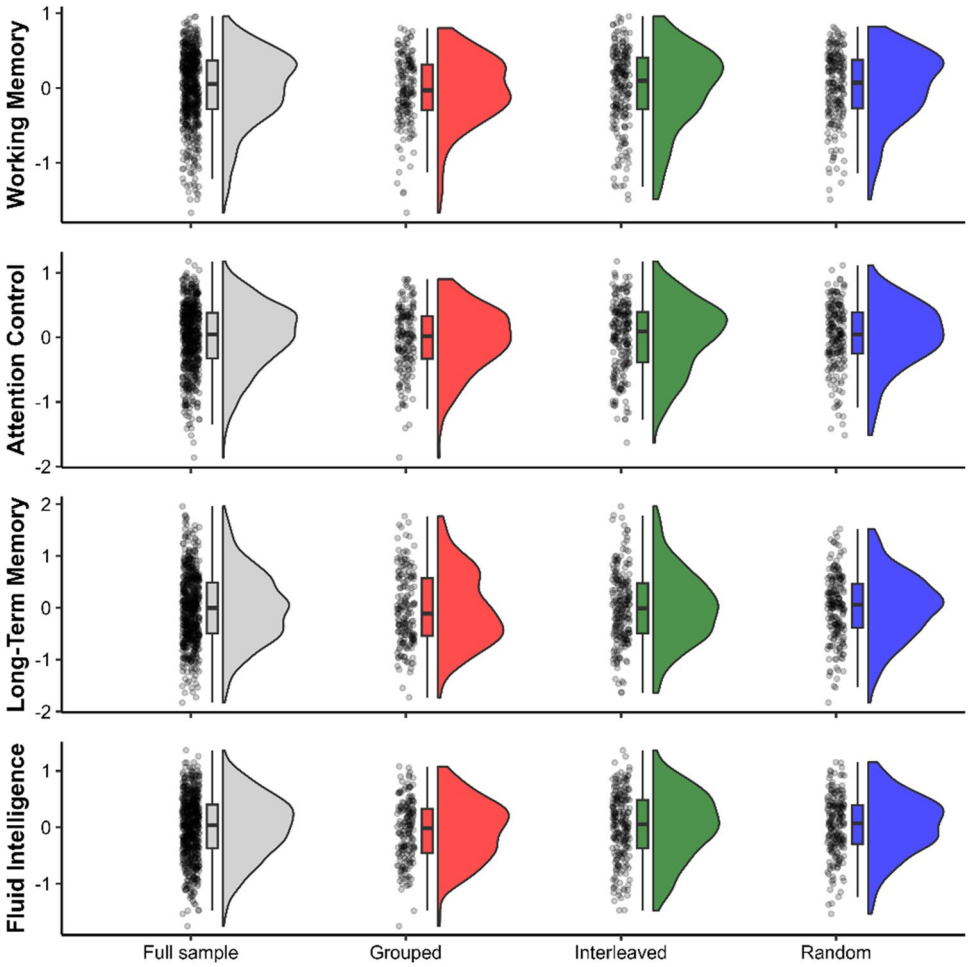

Fig. 5 Raincloud plots of distributions of factor scores by condition (see online article for a color version of this figure).

to situations in which measures of specific constructs are measured on different days of administration. This could be a future extension of the present study. Second, the scope of cognitive abilities measured was also narrow. Therefore, future work may need to perform similar assessments on related cognitive abilities like processing speed, crystallized intelligence, creativity, and problem-solving. Third, we specified our sample sized based on what we estimated would be sufficient to estimate latent factors for a single group ($N = 200$). The power of measurement invariance tests is affected by several factors including sample size, task/item communality, and factor determination (Meade & Bauer, 2007). Some of our tasks, particularly the picture source-recognition task, had relatively low factor loadings (likely due to the programming error which only correctly scored "new" items). Therefore, higher sample size and greater interrelatedness of measures within a construct are more likely to yield measurement invariance. Finally, the study was conducted entirely in a university sample, albeit a diverse one. Future work may need to test for these invariances more systematically with larger and more diverse (i.e., a blend of university and community) samples, more highly correlated manifest variables, and a larger array of cognitive factors.

Ultimately, the present study yields an important, unanswered question: which task sequence is best? Grouped, interleaved, or random? Although we did not observe a systematic effect of sequencing on the factor structure, we would recommend an interleaved task sequence. True randomization is difficult and imposes an administrative burden on the researcher. Fixed task sequencing provides the added benefit of exposing all participants to the same experimental conditions, as argued by Miyake et al. (2000). Unlike experimental approaches, which typically seek to *minimize* between-subject variability outside the specific manipulation and avoid systematic confounds (such as time), the individual-differences approach seeks to *maximize* interindividual variability while minimizing the variability with which participants experience the tasks. Therefore, the individual-differences research seeks to reduce any sources of noise in the measurement that are not "true" *interindividual*

variance in the measures (e.g., task order, time-of-day, light/sounds conditions, task strategies). Therefore, one potential source of noise—the time at which a participant completes a task within a session – can be controlled by fixing the task sequence for all participants. The interleaved design thus represents a nice balance between pragmatics and precision. Regardless, the latent variable approach might be resistant to sequencing effects specifically because it models out measurement noise and estimates latent factors via systematic covariance among putative measurements of a construct.

Conclusions

Variations in task sequences for latent variable analyses of cognitive abilities do not systematically affect average performance or latent factor structures. Although task sequencing did not have a systematic effect here, we recommend a best practice of fixing and interleaving measures of respective constructs to reduce systematic interindividual noise and maximize the likelihood of observing true interindividual variation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13423-023-02369-0.

Funding Authors MKR, GAB, and XC were supported by U. S. Army Research Institute (award No. W911NF2310300).

References

Ackerman, P. L., Beier, M. E., & Boyle, M. D. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General, 131*(4), 567–589.

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., van Langen, J., & Kievit, R. A. (2021). Raincloud plots: A multi-platform tool for robust data visualization [version 2; peer review: 2 approved]. *Wellcome Open Research, 4*(63). https://doi.org/10.12688/wellcomeopenres.15191.2

Aust, F. (2023). Friends don't let friends copy and paste: Reproducible, APA-compliant manuscripts with the R package papaja. *PTOS*, (*2022*). https://doi.org/10.23668/psycharchives.12356

Aust, F., & Barth, M. (2018). *Papaja: Create APA manuscripts with R markdown* [computer software]. Retrieved March 22, 2023, from https://github.com/crsh/papaja.

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language, 66*(3), 407–415.

Buchanan, E. M., Foreman, R. E., Johnson, B. N., Pavlacic, J. M., Swadley, R. L., & Schulenberg, S. E. (2018). Does the delivery matter? Examining randomization at the item level. *Behaviormetrika, 45*(2), 295–316.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. *Cambridge University Press.* https://doi.org/10.1017/CBO9780511571312

Chuderski, A., & Necka, E. (2012). The contribution of working memory to fluid reasoning: Capacity, control, or both? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(6), 1689–1710.

Colom, R., Abad, F. J., Rebollo, I., & Shih, P. C. (2005). Memory span and general intelligence: A latent-variable approach. *Intelligence, 33*(6), 623–642.

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence, 32*(3), 277–296.

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163–183.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51*(1), 42–100.

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers, 17*(6), 652–655.

Dowle, M., & Srinivasan, A. (2021). *Package 'data.Table: Extension of 'data.frame'* (R package version 1.14.2) [computer software]. Retrieved March 22, 2023, from https://CRAN.R-project.org/package=data.table

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2021). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General, 150*(2), 242–275.

Ekstrom, R. B., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests, 1976*. Educational Testing Service.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309–331.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Frith, E., Kane, M. J., Welhaf, M. S., Christensen, A. P., Silvia, P. J., & Beaty, R. E. (2021). Keeping creativity under control: Contributions of attention control and fluid intelligence to divergent thinking. *Creativity Research Journal, 33*(2), 138–157.

Goodhue, D. L., & Loiacono, E. T. (2002). Randomizing survey question order vs. grouping questions by construct: An empirical test of the impact on apparent reliabilities and links to related constructs. Paper presented at the 35th annual Hawaii international conference on system sciences, Big Island, HI.

Hutchison, K. A. (2007). Attentional control and the relatedness proportion effect in semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 645–662.

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130*(2), 169–183.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*(2), 189–217.

Kane, M. J., Meier, M. E., Smeekens, B. A., Gross, G. M., Chun, C. A., Silvia, P. J., & Kwapil, T. R. (2016). Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General, 145*(8), 1017–1048.

Kassambara, A. (2020). *Rstatix: Pipe-friendly framework for basic statistical tests* (R package version 0.6. 0) [computer software].

Kretzschmar, A., & Gignac, G. E. (2019). At what sample size do latent variable correlations stabilize? *Journal of Research in Personality, 80*, 17–22.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*(4), 389–433.

Loiacono, E., & Wilson, V. (2020). Do we truly sacrifice truth for simplicity: Comparing complete individual randomization and semi-randomized approaches to survey administration. *AIS transactions on. Human-Computer Interaction, 12*(2), 45–69.

McGrew, K. S. (2009). *CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research., 37*, 1–10.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling, 14*(4), 611–635.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General, 130*(4), 621–640.

Oberauer, K., Süß, H. M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—Facets of a cognitive ability construct. *Personality and Individual Differences, 29*(6), 1017–1045.

Raven, J. C., & Court, J. H. (1962). *Advanced progressive matrices*. HK Lewis London.

Redick, T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., Kane, M. J., Hambrick, D. Z., & Engle, R. W. (2016). Cognitive predictors of a common multitasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of Experimental Psychology: General, 145*(11), 1473–1492.

Rey-Mermet, A., Gade, M., Souza, A. S., Von Bastian, C. C., & Oberauer, K. (2019). Is executive control related to working memory capacity and fluid intelligence? *Journal of Experimental Psychology: General, 148*(8), 1335–1372.

Richmond, L. L., Burnett, L. K., Morrison, A. B., & Ball, B. H. (2021). Performance on the processing portion of complex working memory span tasks is related to working memory capacity estimates. *Behavior Research Methods, 1–15*. https://doi.org/10.3758/s13428-021-01645-y

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). 'Oops!': Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia, 35*(6), 747–758.

Robison, M. K., & Brewer, G. A. (2020). Individual differences in working memory capacity and the regulation of arousal. *Attention, Perception, & Psychophysics, 82*(7), 3273–3290.

Robison, M. K., & Brewer, G. A. (2022). Individual differences in working memory capacity, attention control, fluid intelligence, and pupillary measures of arousal. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* https://doi.org/10.1037/xlm0001125

Robison, M. K., Gath, K. I., & Unsworth, N. (2017). The neurotic wandering mind: An individual differences investigation of neuroticism, mind-wandering, and executive control. *The Quarterly Journal of Experimental Psychology, 70*(4), 649–663.

Robison, M. K., Miller, A. L., & Unsworth, N. (2020). A multifaceted approach to understanding individual differences in mind-wandering. *Cognition, 198*(*104078*). https://doi.org/10.1016/j.cognition.2019.104078

Robison, M. K., & Unsworth, N. (2018). Cognitive and contextual correlates of spontaneous and deliberate mind-wandering. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(1), 85–98.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Schell, K. L., & Oswald, F. L. (2013). Item grouping and item randomization in personality measurement. *Personality and Individual Differences, 55*(3), 317–321.

Shelton, J. T., Elliott, E. M., Hill, B. D., Calamia, M. R., & Gouvier, W. D. (2009). A comparison of laboratory and clinical working memory tests and their prediction of fluid intelligence. *Intelligence, 37*(3), 283–293.

Shelton, J. T., Elliott, E. M., Matthews, R. A., Hill, B. D., & Gouvier, W. M. (2010). The relationships of working memory, secondary memory, and general fluid intelligence: Working memory is special. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 813–820.

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2015). Working memory capacity and the scope and control of attention. *Attention, Perception, & Psychophysics, 77*(6), 1863–1880.

Shipstead, Z., Lindsey, D. R. B., Marshall, R. L., & Engle, R. W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language, 72*, 116–141.

Thurstone, L. L. (1938). *Primary mental abilities* Psychometric monographs.

Tsukahara, J. S., Harrison, T. L., Draheim, C., Martin, J. D., & Engle, R. W. (2020). Attention control: The missing link between sensory discrimination and intelligence. *Attention, Perception, & Psychophysics, 82*(7), 3445–3478.

Unsworth, N. (2010). On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. *Acta Psychologica, 134*(1), 16–28.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2009a). There's more to the working memory capacity—Fluid intelligence relationship than just secondary memory. *Psychonomic Bulletin & Review, 16*(5), 931–937.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2012). Variation in cognitive failures: An individual differences investigation of everyday attention and memory failures. *Journal of Memory and Language, 67*(1), 1–16.

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology, 71*, 1–26.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498–505.

Unsworth, N., & McMillan, B. D. (2014). Similarities and differences between mind-wandering and external distraction: A latent variable analysis of lapses of attention and their relation to cognitive abilities. *Acta Psychologica, 150*, 14–25.

Unsworth, N., & McMillan, B. D. (2017). Attentional disengagements in educational contexts: A diary investigation of everyday mind-wandering and distraction. *Cognitive Research: Principles and Implications, 2*(1), 1–20.

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009b). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory, 17*(6), 635–654.

Unsworth, N., Redick, T. S., Lakey, C. E., & Young, D. L. (2010). Lapses in sustained attention and their relation to executive control and fluid abilities: An individual differences investigation. *Intelligence, 38*(1), 111–122.

Unsworth, N., Robison, M. K., & Miller, A. L. (2019). Individual differences in baseline oculometrics: Examining variation in baseline pupil diameter, spontaneous eye blink rate, and fixation stability. *Cognitive, Affective, & Behavioral Neuroscience, 19*(4), 1074–1093.

Unsworth, N., Robison, M. K., & Miller, A. L. (2021). Individual differences in lapses of attention: A latent variable analysis. *Journal of Experimental Psychology: General, 150*(7), 1303–1331.

Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory and Language, 62*(4), 392–406.

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2009c). Examining the relations among working memory capacity, attention control, and fluid intelligence from a dual-component framework. *Psychology Science Quarterly, 51*(4), 388–402.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. *Springer.* https://doi.org/10.1007/978-0-387-98141-3.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pederson, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. 10.21105/joss.01686.

Wilke, C. O., Wickham, H., & Wilke, M. C. O. (2019). *Package 'cowplot': Streamlined plot theme and plot annotations for 'ggplot2'* [computer software]. Retrieved March 22, 2023, from https://CRAN.R-project.org/package=cowplot.

Wilkinson, R. T., & Houghton, D. (1982). Field test of arousal: A portable reaction timer with data storage. *Human Factors, 24*(4), 487–493.

Wilson, E. V., & Lankton, N. (2012). Some unfortunate consequences of non-randomized, grouped-item survey administration in IS research. *Proceedings of the 33rd International Conference on Information Systems.*

Wilson, E. V., Srite, M., & Loiacono, E. (2017). *A call for item-ordering transparency in online IS survey administration* Proceedings of Americas Conference on Information Systems.

Wilson, V., Srite, M., & Loiacono, E. (2021). The effects of item ordering on reproducibility in information systems online survey research. *Communications of the Association for Information Systems, 49*(1), 41.

Open practices statement All data, analysis code (in R), and task scripts (in Python) are available on the Open Science Framework (https://osf.io/a79hf/). This study was not preregistered.