NOTE: This is a non-peer reviewed preprint

From task-general to task-specific in a few minutes: Working memory performance is an adaptive process

Running head: From general to specific

Jussi Jylkkä[a], Daniel Fellman[a,b,c], Otto Waris[d,e], Liisa Ritakallio[a], Juha Salmi[f,g,h], Matti Laine[a]

[a]Department of Psychology, Åbo Akademi University, Finland

[b]Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden

[c]Department of Applied Educational Science, Umeå University, Umeå, Sweden

[d]Department of Child Psychiatry, University of Turku and Turku University Hospital, Finland

[e]INVEST Research Flagship Center, University of Turku, Turku, Finland

[f]Department of Neuroscience and Biomedical Engineering, Aalto University, Finland

[g]Department of Psychology and Speech-Language Pathology, University of Turku, Finland

[h]Turku Institute for Advanced Studies, University of Turku, Finland

Corresponding author: Jussi Jylkkä, jjylkka@abo.fi, Department of Psychology, Åbo Akademi University, Fabriksgatan 2, 20500 Åbo, Finland

Declarations of interest: none

Acknowledgments: OW was funded by the Academy of Finland INVEST Flagship Programme (#320162); JS was supported by the Academy of Finland (#325981 and #328954); ML was supported by Academy of Finland (#323251).

Abstract

Measurement of cognitive functions is often based on the untested assumption that the mental architecture underlying cognitive task performance is constant throughout the task. In contrast, skill learning theory implies that any task performance is an adaptive process that progresses from effortful and task-general to automatic and task-specific. This hypothesis has not been applied to the short time spans of traditional cognitive tasks such as working memory (WM) tasks. We ran consecutive confirmatory factor analyses on two well-powered data sets to test the hypothesis that the initial stages of WM task performance are more strongly related to task-general mechanisms (a single g-factor) whereas at later stages, task-specific dissociable skills gain prominence. In both experiments (N = 296 and N = 201), the g-model showed good fit to the data in the very first stages of task performance, but its fit steadily declined over the task period. The fit of the task-specific model was adequate throughout task performance and outperformed the g-model shortly after task onset. The results suggest that the mental architecture underlying complex cognitive performance changes rapidly from initial engagement of task-general metacognitive and control processes to the predominance of task-specific skills. This has several important implications for the measurement of complex cognition.

Keywords: Working memory, skill learning, latent structure, factor analysis, cognitive architecture, construct validity, reliability, task performance

1. Introduction

Survival requires adaptation. According to Spencer (1855), the "fundamental condition of vitality is that the internal order shall be continually adjusted to the external order" (§173). On this line of reasoning, cognition is an adaptive process. Novel tasks call for adaptive behavior that is thought to rely on successful executive control (Miller & Cohen, 2001; Miyake et al., 2000; Norman & Shallice, 1986). While adaptivity and learning are considered to be cornerstones of human cognitive ability (Anderson, 2014; Chein & Schneider, 2012), the outcome measures of complex cognitive task performances typically sum up the whole task period into single scores. This leads to a static view of the mental architecture underlying task performance. Anderson (2014) calls this the *componential computational theory of mind*, which implies that the mind consists of dissociable functions that can be measured with cognitive tasks. For example, the Stroop task "taps on" a person's inhibitory capacity and the n-back task taps on their working memory updating capacity. Considering cognition as an adaptive process is a radical departure from this approach: it entails that performing a cognitive task changes what is being measured, because the cognitive system adapts to the context. Thus, what is measured by a task is partly created by the measurement situation, rather than being "out there" on the latent level, waiting to be measured. This way of thinking motivates examining cognitive performance as a dynamic process that evolves over time.

The skill learning approach (Chein & Schneider, 2012; Schneider & Chein, 2003; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977) espouses this general view, by focusing on the adaptivity of cognition instead of static mental architecture. It is based on a long research tradition on slow and effortful vs. fast and automatic cognitive systems. The skill learning approach is often applied to explain learning in perceptual-motor tasks or tasks that require problem solving. In the present study, we apply it to complex cognitive task performance, namely working memory (WM) tasks. These tasks involve perceptual-motor processes, and while they do not require problem solving, the cognitive demands and unfamiliarity of the tasks call for executive resources needed for strategy generation and implementation (e.g., Waris, Fellman, Jylkkä, & Laine, 2020; Waris, Jylkkä, Fellman, & Laine, 2021). In this way, they can be considered as analogical to any novel task that requires skill to be performed efficiently. The skill learning approach entails that processing of novel

tasks proceeds from task-general metacognitive and control processes to task-specific, more automatic processes. Metacognitive processes include the generation of a suitable strategy for the task while the controlled execution phase involves its effortful and controlled implementation. Automatization proceeds gradually with task exposure and frees the effortful processes to other tasks (Chein & Schneider, 2012).

Applying the skill learning approach to the psychometric context implies that what is being measured can change as the participant develops a skill to perform the task. Consider a complex WM task like the n-back, where the subject sees stimuli that are constantly changing. During each new stimulus, the participant needs to decide whether the stimulus is the same as the one that was presented *n* trials ago. For example, in a 2-back task with the series 2-5-1-**5**-4-3-**4**, the subject should press the response button corresponding to "same" during the bolded trials (Kirchner, 1958). The participant assumedly enters the metacognitive phase already upon receiving task instructions (wondering what the task will be like, trying to remember the stimulus-response mappings and task instructions, planning a strategy, etc.). Thus, cognitive processes responsible for task performance start even before the first trial. The strategies and memorized task instructions are effortfully implemented in the beginning of a task, but processing is gradually automatized. The relative contribution of executive control is expected to diminish, while the task-specific skills continue to become routinized. Thus, as the cognitive processes engaged by the task evolve in time, it can be questioned whether a cognitive task relies on a fixed constellation of cognitive processes, as is commonly assumed. Following the skill learning view (Chein & Schneider, 2012), we hypothesize that a cognitive task relies on different processes depending on the stage of learning, and that such changes can take place already during the brief time spans (minutes) that these tasks usually take.

The hypothesis that the later stages of task performance rely on task-specific automatizations receives support from the cognitive training literature, albeit there the time scales are hours or days, rather than minutes. Prolonged training on a cognitive task typically does not yield transfer to other types of tasks, but rather enhances performance only in tasks that have close structural resemblance to the trained task (Au et al., 2015; Karbach & Verhaeghen, 2014; Kassai, Futo, Demetrovics, & Takacs, 2019; Melby-Lervåg et al., 2016; Melby-Lervåg & Hulme, 2013; Sala & Gobet, 2017; Schwaighofer, Fischer, & Bühner, 2015; Weicker, Villringer, & Thöne-Otto, 2016). For example, n-back training with digit stimuli can clearly enhance performance on a untrained n-back task with letters, but not on structurally dissimilar WM updating tasks such as a running memory task (Gathercole, Dunning, Holmes, & Norris, 2019; Soveri, Antfolk, Karlsson, Salo, & Laine, 2017). This lack of transfer can be

explained if we consider cognition as an adaptive process, where performance on a repeated task starts to automatize and rely on task-specific subroutines instead of task-general processes. Also, the brain activation findings in WM training, showing modulations of activity in dorso- and ventrolateral prefrontal cortex and perceptual-motor systems before vs. after weeks of training, have been interpreted in the skill learning context (Salmi, Nyberg, & Laine, 2018).

In the present study, we examined for the first time the evolvement of latent structure of commonly used WM tasks within a single testing session taking ca 5 to 14 minutes per task. This is motivated by neuroimaging evidence indicating that the brain systems underlying WM begin to adapt already during one testing session (e.g., Badre, Kayser, & D'Esposito, 2010). Another line of evidence for the dynamic nature of the cognitive processes underlying within-session WM performance comes from self-reported strategies that are also reliably associated with objective WM performance (Fellman et al., 2020; Forsberg, Fellman, Laine, Johnson, & Logie, 2020; Laine, Fellman, Waris, & Nyman, 2018). Of particular relevance here is a very recent microgenetic study (Waris et al., 2021) that analyzed the block-by-block evolvement of strategy use within a single n-back test session. About half of the participants reported using a strategy already during the very first task block, and changes in selected strategy were most common during the initial task blocks, after which strategy use became more stable. The generality of this pattern in strategy development was corroborated by a similar study that successfully replicated these findings with quite different episodic memory tasks (Waris et al., 2020). The findings of these two microgenetic studies fit well to the skill learning view, according to which strategy generation and controlled execution are most prominent at the initial stages of task performance. Importantly, they indicate that changes in the allocation of cognitive resources do not require a longer practice period, such as in WM training, but can be observed even in a single test session, i.e., within a few minutes. Nevertheless, these results are only suggestive as they are derived from subjective strategy reports.

1.1. Earlier empirical research on the latent structure of WM performances

The latent structure of WM has been addressed in several factor analytic studies, but the results have been mixed. Some studies indicate that WM is primarily organized by its contents (numerical-verbal vs. spatial; Mackintosh & Bennett, 2003; Shah & Miyake, 1996; Vuong & Martin, 2014; Waris et al., 2017), while other studies have suggested divisions according to WM sub-processes (e.g., maintenance, updating; Bledowski, Rahm, & Rowe,

2009; Miyake et al., 2000; Unsworth, Fukuda, Awh, & Vogel, 2014) or obtained a single general latent WM factor (e.g., Colom, Shih, Flores-Mendoza, & Quiroga, 2006; Wilhelm, Hildebrandt, & Oberauer, 2013). These discrepancies may relate to differences in the WM test batteries, samples, and the utilized analytical approach, but possibly also in part to the assumption that WM task performances reflect a static underlying cognitive structure. If WM task performance is an adaptive process as the skill learning approach entails, the cognitive layout seen in factor analyses may also reflect individual variation in what stage of learning a person is in a task: in the beginning of a task one is likely to utilize task-general processes, whereas in the latter part one may rely on more task-specific processes. Thus, depending on the difficulty and length of the WM tasks as well as participant characteristics, the participants may find themselves at different stages of their skill learning processes, which would increase variance in factor analytic findings between different studies.

What is lacking in these earlier studies of the latent structure of WM processes is an analysis that takes time into account, by examining how the latent structure changes during the brief period of WM task performance. We addressed this question in the present dual-experiment study. Although we focused on WM tasks, it is worth noting that the skill learning approach tested here would make the same predictions about any complex cognitive tasks, not only WM.

1.2. The present study

To test the assumption that the cognitive processes underlying WM task performance are not static but change during the short time period such tasks usually take, we analyzed the latent structure of two independent data sets (WM task battery administered to healthy adults) by running sequential confirmatory factor analyses (CFA) on the initial, mid, and final sections of task performance. Our data consisted of pretest results from two previous WM training experiments (the first one reported in Fellman et al., 2020, the other hitherto unpublished preregistered at https://osf.io/c9ygt). Unlike these two previous experiments where the present data stems from, the current study is not preregistered. The WM task paradigms employed as pretest measures were n-back, running memory, simple span, and selective updating. Each WM paradigm included task variants that employed different stimuli.

In the CFAs, we used two types of models: a single g-factor model and a paradigm-specific model (three or four WM task paradigms, each with two to four task variants). We pitted two hypotheses against each other. The *adaptive cognition hypothesis* that we embraced states that processing of novel tasks evolves from more task-general, global, and united to

more task-specific, modular, and diverse. This was operationalized in CFA as follows: in the initial performance of complex, unfamiliar WM tasks, the fit of a general g-factor model is better than that of a paradigm-specific factor model, but the fit of the g-factor model declines over time as the tasks start to become routinized. We also expected the fit of the paradigm-specific model to increase over time, reflecting the development of task-specific subroutines or automatizations. In contrast, the *static cognition hypothesis* that is at least implicitly dominant in the whole field of cognitive testing, holds that the cognitive architecture and thus the latent factor structure of our complex WM tasks remains constant throughout the task session.

2. Experiment 1

In this experiment, we tested the two main hypotheses with data from an Internet-based WM training study. In order to examine the underlying cognitive architecture of WM processing when the tasks are novel, we focused solely on pretest performances. The duration of the tasks was comparable to that used in cognitive testing generally (on average 5 – 14 minutes each).

2.1 Participants and procedure

The current data stems from the pretest of an online randomized controlled trial that examined WM training and its mechanisms (Fellman et al., 2020; preregistration at https://aspredicted.org/r7qs9.pdf). The study was approved by the Institutional Review Board of the Departments of Psychology and Logopedics, Åbo Akademi University, and it was conducted in accordance with the Helsinki Declaration. In that study, 419 participants were recruited through Prolific Academic (https://www.prolific.co/). Of these, 296 participants were included in the present study after excluding those that did not met our inclusion criteria (n = 117), and those that reported cheating during task performance (n = 6). Our inclusion criteria were that participants are English native speakers, have no current psychiatric or neurological illnesses that affected daily life, no current use of central nervous system (CNS) medication, and no current psychotropic drug use (except tobacco, alcohol, and cannabis). The participants were 18–50-year-old with a mean age of 34.08 years (SD = 8.52), and 62.50 % (n = 185) were females. The mean length of education in the sample was 15.80 years (SD = 3.49).

2.2. Materials

The eight WM tasks included in this study encompassed four n-back tasks (separate tasks with digits, letters, colors, and boxes as stimuli), four Simple span tasks (Forward simple spans with digits and boxes, Backward simple spans with digits and boxes), and two Running memory tasks (digits and boxes). All data were from the pretest of the original experiment when the participants saw these tasks for the first time (Fellman et al., 2020).

2.2.1 N-back

We employed four adaptive single n-back tasks: with digits (1 to 9), letters (A to I), colors (blue, yellow, red, green, purple, black, pink, orange, and gray), and locations (boxes presented in a 3 × 3 matrix) as stimuli. The items in each n-back task variant were presented one at a time on a computer screen, and the participants were instructed to respond "yes" or "no" to each item with a computer keyboard press, indicating whether or not the current item corresponded to the item presented n items back in the sequence. Each n-back variant consisted of 12 blocks of n-back, with each block containing 20 + *n* trials. Out of the 20 trials in a block, six were targets and 14 non-targets. Four of the non-targets were lures (i.e., identical to the target items except that they were presented n ± 1 back) to increase the task demands. Stimulus display time for each item in a sequence was 1500 milliseconds, whereas the interstimulus interval was 450 milliseconds. The n-back tasks were adaptive so that task difficulty depended on individual success rate. Each n-back task started with a 1-back block, and the level of *n* could vary between 1 and 12. If the participant recalled 18–20 trials correctly in a block, the program increased *n* by one. The level of *n* remained the same if the participant recalled 15–17 trials correctly, while 5 or more incorrectly recalled trials resulted in a decrease of *n* by one. The length of these tasks was ca 14 minutes each.

2.2.2. Span

We administered four simple span tasks: two Forward span tasks and two Backward span tasks. One of the forward spans used digits as stimuli (Forward simple span with digits; FSD), whereas the other employed visuospatial locations (Forward simple span with boxes; FSB). In both variants, the participants completed one trial of each list length, with the length of the randomized lists ranging from 4-10 in FSD and 3-9 in the FSB. We also administered two Backward simple span tasks, one with digits (Backward simple span with digits; BSD), and the other one with locations (Backward simple span with boxes; BSB). The list lengths ranged from 3 to 9 in both Backward span variants, and the item sequences were presented in

a randomized order. As in the Forward simple span tasks, the participants were to recall one sequence per list length. Both the Forward and Backward simple span task had the same stimulus presentation time, with a 1000 millisecond item exposure time and 500 millisecond inter-stimulus intervals. The length of these tasks was ca 5 minutes each.

2.2.3. Running memory

The test battery included two Running memory tasks where the participants were prompted to recall a given number of last items from a suddenly aborted sequence. One of those was a Running memory task with digits (RMD) whereas the other one was a Running memory task with boxes (RMB). In both variants, stimulus presentation time was 1000 milliseconds and interstimulus interval was 500 milliseconds. In RMD, digit sequences with unpredictable length were shown on the screen, after which the participant was to recall the last four items in the correct order. The RMB was otherwise identical to the RMD, but the stimuli were spatial locations presented in a 3 × 3 matrix. In both task variants, the participant completed eight trials in a randomized order, with item sequences ranging from 4 to 11 items (i.e., one trial of each list length). The length of these tasks was ca 6 minutes each.

2.3. Analysis plan in Experiment 1

To test the adaptive vs. static hypothesis, we used confirmatory factor analyses in two ways. First, to get an overall view of the WM latent structure with the present task battery, we conducted CFAs in the way they have thus far been used in WM research, i.e., by using summative data (final scores) from the tasks. Second, we examined whether the factor structure changed during task performance by running consecutive CFAs where the data was evenly split into three consecutive "time slices" based on trial number. We call these full data analysis and three-split analysis, respectively. A g-factor model, consisting of all ten WM tasks, served as our baseline model. This model was compared against a three-factor paradigm-specific model, consisting of an n-back factor (NBD, NBL, NBC, NBB), a Running memory factor (RMD and RMB), and a Span factor (FSD, FSB, BSD, BSB) (see Figure 1). The paradigm-specific model was chosen as the reference model to the g-factor model, as the results of many WM training studies have highlighted the importance of the task paradigm: repeated practice leads to paradigm-specific learning, seen as transfer to unpracticed variants of the trained task (Au et al., 2015; Karbach & Verhaeghen, 2014; Kassai, Futo, Demetrovics, & Takacs, 2019; Melby-Lervåg et al., 2016; Melby-Lervåg & Hulme, 2013; Sala & Gobet, 2017; Schwaighofer, Fischer, & Bühner, 2015; Weicker, Villringer, & Thöne-Otto, 2016).

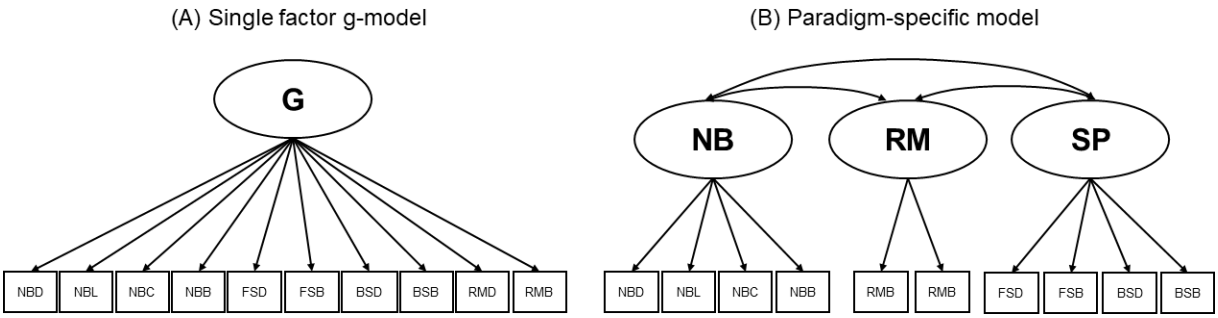

Figure 1. Candidate models for the analyses in Experiment 1. G = g-factor; NB = N-back; RM = Running memory; SP = Span; NBD = N-back with digits; NBL = N-back with letters; NBC = N-back with colors; NBB = N-back with boxes; RMD = Running memory with digits; RMB = Running memory with boxes; FSD = Forward span with digits; BSD = Backward span with digits; FSB = Forward span with boxes; BSB = Backward span with boxes.

2.4. Dependent variables

We first conducted CFAs on the whole dataset in the traditional way by using summative data. For the n-back tasks, the dependent variables were the average level of n achieved across the 12 n-back blocks. For the Running memory and the span tasks, the dependent variables were the total number of correctly recalled items in their correct position across all trials.

Next, to examine whether the latent structure changed over time, we split the sequences into three phases (initial, mid, final), roughly matched in length. The divisions were as follows (amount of blocks at initial/mid/final): 4/4/4 in the n-back tasks (i.e., NBD, NBL, NBC, NBB); 2/2/3 in the Running memory tasks (i.e., RMD, RMB) ; and 2/2/3 in the Span tasks (i.e., FSD, FSB, BSD, BSB). Thus, where an even split was impossible, the remaining blocks were included in the last phase. This is motivated by the assumption that most changes in cognitive processing happen at the initial stages of task performance (Chein & Schneider, 2012). The dependent variable for the n-back tasks consisted of the highest n-back level achieved in the respective task phase. For the Running memory tasks, the dependent variables were calculated by counting the correctly recalled items in correct position within the respective task phases. For the Span tasks, the dependent variables consisted of the proportion

of correctly recalled items in correct serial order within the respective task phases (i.e., number of correctly recalled items/total number of items).[1]

2.5. Model fit and comparison

Models were estimated in the lavaan software package (Version 0.5-20; Rosseel, 2012) in R version 3.1.3 (R Core Team, 2015) using maximum likelihood estimation and robust standard errors. Missing observations were dealt with the full maximum likelihood (FIML) parameter estimation technique. The overall fit of each model was assessed using the χ^2 test, the comparative fit index (CFI), the root mean square error of approximation (RMSEA) which is reported with 90% confidence intervals, the standardized root mean square residual (SRMR), and the Akaike information criterion (AIC). For CFI, values larger than .95 indicate good fit, and values between .90 and .95 indicate acceptable fit. RMSEA values smaller than .06 and SRMR values smaller than .08 indicate a good fit (Hu & Bentler, 1999). Single g-factor and paradigm-specific models from the corresponding task phase which were executed without any convergence problems were further compared via a likelihood ratio test (i.e., the scaled χ^2 difference test). AIC is a modified version that takes into consideration the ‘‘complexity’’ of the evaluated model (in terms of degrees of freedom) and penalizes more complex models (i.e., models with fewer degrees of freedom). AIC values are generally used when comparing non-nested or non-hierarchical models estimated with the same data and indicates to the researcher which of the models is the most parsimonious one (Hooper, Coughlan, & Mullen, 2008). Smaller AIC values suggest a good fit. As our two proposed models were nested, their fit comparisons were evaluated with the likelihood ratio test, but we nevertheless report AIC values for the sake of transparency.

2.6. Results of Experiment 1

The data were screened to identify univariate outliers (i.e. scores deviating >3.5 *SD*s from the sample mean on each task). Moreover, two participants reported being color-blind: performances from these participants in the n-back color task were thus removed. These criteria together with empty cells due to technical problems during data collections resulted to 0.61% ($n_{observations}$ = 18) of missing data in the full-data analysis and 0.48% ($n_{observations}$ = 43)

[1] A proportion score was calculated instead of a summative score due to the fact that the sequences, which varied in length (i.e., participants had the possibility to receive more correctly recalled items in lengthier sequences), were randomized across participants.

of missing data in the three-split analysis. Descriptive statistics and pairwise correlations between tasks are summarized in Supplementary Materials, Appendix A.

2.6.1. Full-data analysis

The results from the CFAs on the full dataset (see Table 1) showed that the g-factor solution possessed poor fit (χ^2 (35) = 233.458, CFI = 0.847, RMSEA = 0.137, (90% CI = 0.119, 0.155), SRMR = 0.080), whereas the paradigm-specific model showed good fit (χ^2 (32) = 100.871, CFI = 0.948, SRMR = 0.047), except in the RMSEA index, which was slightly above the threshold of being acceptable (RMSEA = 0.084). When comparing the g-factor model with the paradigm-specific model using the chi-square difference test, the results showed that the paradigm-specific model exhibited better fit to the data compared to the g-factor model ($\Delta\ \chi^2$ = 81.381, $\Delta\ df$ = 3, $p < 0.001$). Factor loadings of the two models are depicted in Figure 2.

Table 1 Fit statistics for both models included in the confirmatory factor analyses on the full data in Experiment 1.

							Single-factor G vs. Paradigm-specific		
Model	χ^2	*df*	CFI	SRMR	RMSEA	AIC	$\Delta\chi^2$	Δdf	*p*
Single-factor G	233.458	35	0.847	0.080	0.137 [0.119, 0.155]	7263	81.381	3	< .001
Paradigm-specific	100.871	32	0.948	0.047	0.084 [0.065, 0.104]	7136			

Note. $\chi 2$ = Chi-square; *df* = degrees of freedom; Δ = difference; *p* = *p* value; CFI= comparative fit index; RMSEA = root mean square error of approximation (90% confidence intervals are given in square brackets); SRMR= standardized root-mean-square residual; AIC= Akaike information criterion.

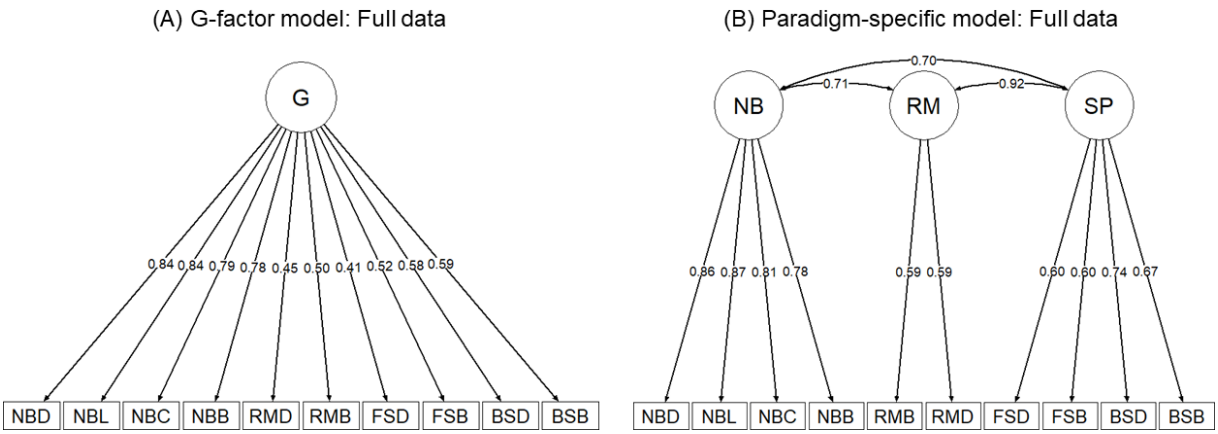

Figure 2. Full data: (A) A single g-factor model where all tasks load on a common factor; (B) A paradigm-specific model. Latent factors are shown in ovals and squares represent observed variables. G = g-factor; NB = N-back; RM = Running memory; SP = Span; NBD = N-back with digits; NBL = N-back with letters; NBC = N-back with colors; NBB = N-back with boxes; RMD = Running memory with digits; RMB = Running memory with boxes; FSD = Forward span with digits; FSB = Forward span with boxes; BSD = Backward span with digits, BSB = Backward span with boxes. All parameter estimates shown are fully standardized.

2.6.2. Three-split analyses

The results from the CFAs at the **initial task phase** in the three-split analyses (see Table 2) showed that the g-factor solution had a good fit (χ^2 (35) = 55.187, CFI = 0.949 RMSEA = 0.043, (90% CI = 0.016, 0.065), SRMR = 0.041), as did the paradigm-specific model (χ^2 (32) = 42.217, CFI = 0.967, RMSEA = 0.038, (90% CI = 0.000, 0.060), SRMR = 0.038). The paradigm-specific model was shown to fit the data better compared with the g-factor model ($\Delta\chi^2$ = 8.02, Δ *df* = 3, *p* = 0.046) (see also Fig. 3A, and 3B depicting the factor loadings of the two models).

With respect to the **mid task phase**, the g-factor model fit to the data was poor-to-acceptable (χ^2 (35) = 82.692, CFI = 0.935, RMSEA = 0.067 (90% CI = 0.048, 0.086), SRMR = 0.053). The paradigm-specific model, in turn, showed a good fit, (χ^2 (35) = 82.692, CFI = 0.986, RMSEA = 0.032 (90% CI = 0.000, 0.057), SRMR = 0.035), and outperformed the g-factor model ($\Delta\chi^2$ = 32.014, Δ *df* = 3, *p* < 0.001). However, as depicted in Figure 3D, the factor correlation between Running memory and Span in the paradigm-specific model was inadmissible (i.e., greater than 1), producing a covariance matrix that was not positive definite. This suggests that the manifest variables under these two factors should be merged under one and the same latent factor. Nevertheless, to keep the mid-task phase coherent with the other phases, we refrained from computing additional models.

Regarding the **final task phase**, the g-factor model showed a poor fit to the data (χ^2 (35) = 138.217, CFI = 0.890, RMSEA = 0.100 (90% CI = 0.083, 0.118), SRMR = 0.070),

whereas the paradigm-specific model showed a good fit (χ^2 (32) = 40.223, CFI = 0.999, RMSEA = 0.031, (90% CI = 0.000, 0.056), SRMR = 0.032). When comparing the models, the g-factor model was again outperformed by the paradigm-specific model ($\Delta\ \chi^2$ = 70.706, $\Delta\ df$ = 3, $p < 0.001$) (see also Fig. 3E, and Fig. 3F depicting the factor loadings of the two models). To assess the hypothesis that the fit of the g-factor model would go down and that of the paradigm model would go up with time, fit indices as a function of time are depicted in Figure 4.

Table 2 Fit statistics for both models included in the confirmatory analyses on the three-split data analyses in Experiment 1.

							Single-factor G vs. Paradigm-specific		
Model	χ^2	*df*	CFI	SRMR	RMSEA	AIC	$\Delta\chi^2$	Δdf	*p*
					Initial task phase				
Single-factor G	55.187	35	0.949	0.041	0.043 [0.016, 0.065]	8043	8.02	3	0.046
Paradigm-specific	44.215	32	0.967	0.038	0.036 [0.000, 0.060]	8038			
					Mid task phase				
Single-factor G	82.692	35	0.935	0.053	0.067 [0.048, 0.086]	7714	32.014	3	< .001
Paradigm-specific	42.217	32	0.986	0.035	0.032 [0.000, 0.057]	7679			
					Final task phase				
Single-factor G	138.217	35	0.890	0.070	0.100 [0.083, 0.118]	7427	70.706	3	< .001
Paradigm-specific	40.223	32	0.999	0.032	0.031 [0.000, 0.056]	7335			

Note. χ2 = Chi-square; *df* = degrees of freedom; Δ = difference; *p* = *p* value; CFI= comparative fit index; RMSEA = root mean square error of approximation (90% confidence intervals are given in square brackets); SRMR= standardized root-mean-square residual; AIC= Akaike information criterion.

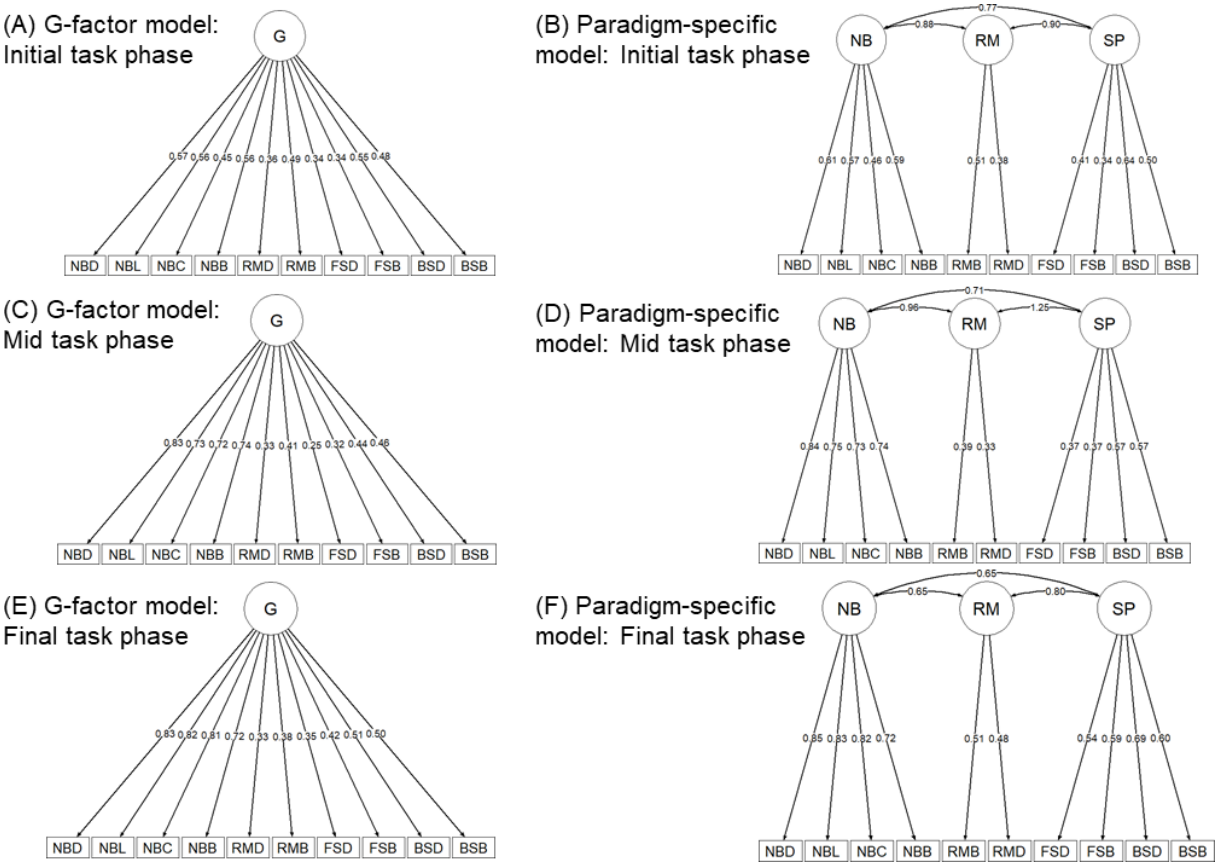

Figure 3. Three-split data: (A) A single g-factor model where all tasks load on a common factor on initial task performance; (B) A paradigm-specific model on initial task performance. (C) A single g-factor model where all tasks load on a common factor on mid task performance. (D) A paradigm-specific model on mid task performance (note that the inadmissible factor correlation between RM and SP produced a covariance matrix that was not positive definite). (E) A single g-factor model where all tasks load on a common factor on final task performance. (F) A paradigm-specific model on final task performance. Latent factors are shown in ovals and squares represent observed variables. G = g-factor; NB = N-back; RM = Running memory; SP = Span; NBD = N-back with digits; NBL = N-back with letters; NBC = N-back with colors; NBB = N-back with boxes; RMD = Running memory with digits; RMB = Running memory with boxes; FSD = Forward span with digits; FSB = Forward span with boxes; BSD = Backward span with digits, BSB = Backward span with boxes. All parameter estimates shown are fully standardized.

2.6.3. Post hoc analyses

In the three-stage analyses, there was a clear trend that the g-factor model fit was good at the initial stage but became worse towards the end of the tasks. At the same time, the paradigm-specific model was near-significantly better already at the initial stage. To look further into this, we ran post hoc analyses on the very initial blocks of the tasks, as it is possible that with the present demanding but as such straightforward WM tasks, the hypothesized initially strong engagement of the Metacognitive and Cognitive control systems is very short-lived. In all tasks except for the n-back, only the very first sequence from the

task was included. In the n-back, the first three blocks were included, because the first two blocks do not show sufficient variance in the dependent variable (average n-back level): the first block is always 1-back and the second block is at highest 2-back, which almost all participants reach. At this very initial stage, which we denote as "pre-initial", the g-factor model (χ^2 (35) = 40.232, CFI = 0.975, RMSEA = 0.020 (90% CI = 0.000, 0.054), SRMR = 0.038), and the paradigm-specific model (χ^2 (32) = 39.482, CFI = 0.956, RMSEA = 0.028 (90% CI = 0.000, 0.0540), SRMR = 0.038) did not differ in model fit ($\Delta\ \chi^2 = 0.574$, $\Delta\ df = 3$, $p = 0.902$), and the fit indices of the g-factor model were slightly better than those of the paradigm-specific model. It is worth pointing out that for the paradigm-specific model, the high correlations between the latent factors ($r = .78 – .99$) resulted to a covariance matrix that was not positive definite. Figure 4 illustrates how the fit of the models evolves over time.

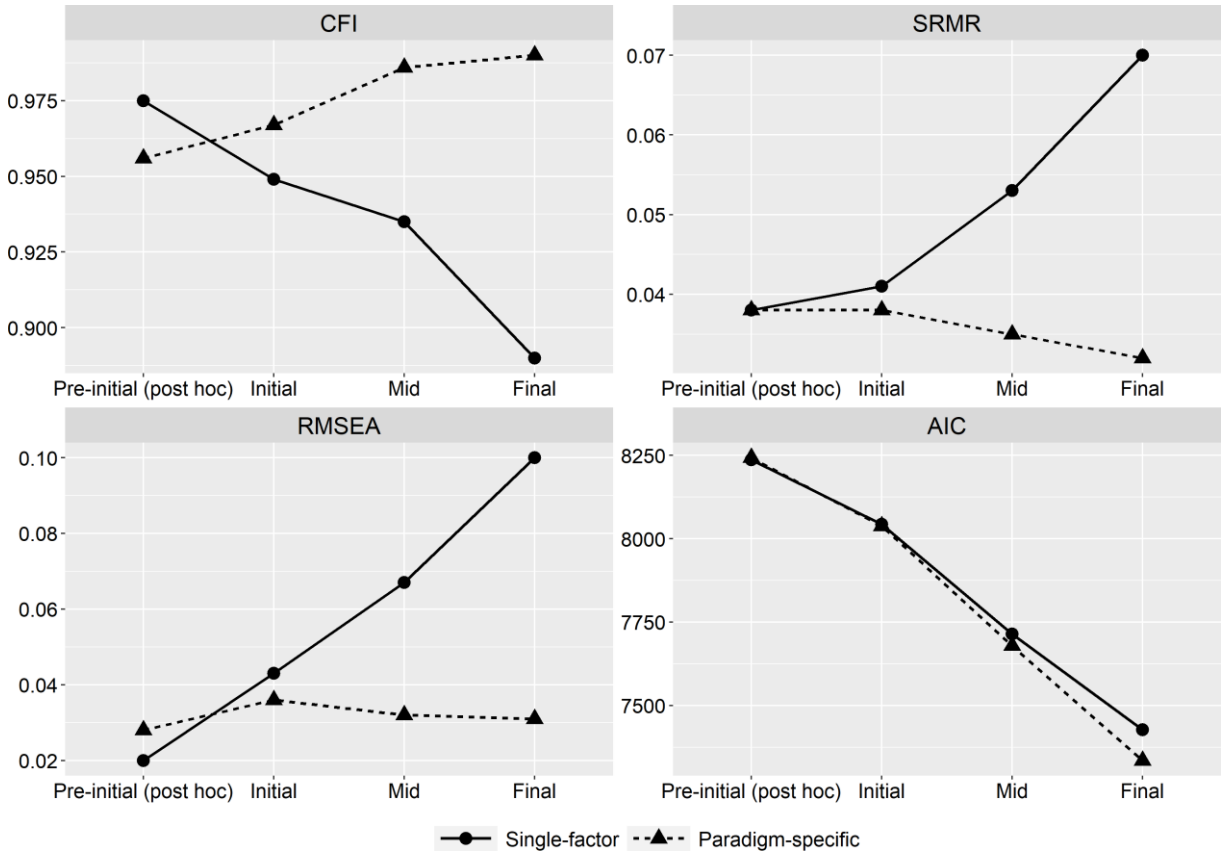

Figure 4. Fit indices of the single g-factor model vs. the paradigm-specific factor model as a function of time (initial, mid, and final) in Experiment 1. CFI= comparative fit index (larger is better); RMSEA = root mean square error of approximation (smaller is better); SRMR= standardized root-mean-square residual (smaller is better); AIC= Akaike information criterion (smaller is better). Note: the paradigm-specific model failed to converge at mid-phase, so those results are not reliable.

2.7. Discussion (Experiment 1)

Skill learning theories imply that processing of novel cognitive tasks proceeds adaptively from task-general to task specific (e.g., Chein & Schneider, 2012). At the initial phases of task performance, metacognitive processes such as strategy generation and cognitive control should be central, whereas at later phases, task performance starts to become at least partly automatized. We analyzed commonly used WM tasks with consecutive CFAs where performance within each brief testing session was divided into initial, mid and final task phases. We assumed that task-general executive control is reflected in the fit of the g-factor model, which we expected to show best fit indices in the beginning of task performances, followed by deteriorating fit indices as the tasks are processed more automatically. Conversely, we expected the fit indices of the paradigm-specific model to become better across task phases, reflecting automatization. As an alternative, we had the static cognition hypothesis which holds that the latent factor structure remains constant within the task sessions.

The results were in line with the adaptive cognition hypothesis with respect to the g-factor model, as its fit steadily declined (see Figure 4). However, the fit of the paradigm-specific model was constantly good throughout task performance and did not improve over time as we expected (we will discuss this finding in more detail in the General discussion). The paradigm-specific model outperformed the g-factor model at all three task phases, contrary to the hypothesis that predicted initial prominence for the g-factor model. This could be because the highest engagement of task-general metacognitive and control processes takes place still earlier. This possibility is supported also by the changes in WM strategy use reported by Waris et al. (2020, 2021), which cluster mainly in the initial 2-3 task blocks. A post hoc CFA targeting this pre-initial stage showed that the g-factor model was not outperformed by the paradigm-specific model, and the fit indices of the g-factor model were slightly better overall, in line with the adaptive hypothesis. It is also noteworthy that in the nested model comparisons that operate on chi-square, a significant result can only indicate that the model with more free parameters (i.e., here the paradigm-specific model) is better able to reproduce the data as compared to the base model (i.e., here the g-factor model), and not vice versa (e.g., Breckler, 1990). In other words, the g-factor model that serves as baseline can never show better fit than the paradigm-specific model.

In sum, the present findings support the hypothesis that the latent structure of WM task performances undergoes a rather rapid development where task-general executive processes are active at the very beginning of task performance, although this phase passes

quickly. These results thus challenge the assumption that complex cognitive tasks tap on static cognitive processes, an assumption that is implicitly built into most cognitive psychology experiments and assessments.

One limitation regarding the interpretation of the results is that in the paradigm-specific model, the Span and Running Memory tasks largely formed a unitary factor (*r*'s ≥ .8), indicating that the paradigm-specific model is effectively formed by an n-back factor together with a shared factor for Span and Running memory. This could be because the n-back task was adaptive unlike the other tasks. However, if the adaptivity of the n-back task would influence the model fit indices, the effect should be contrary to what we hypothesized, since an adaptive task should exert a higher executive load throughout the task, as the participant is kept working on upper performance limits.

Due to the theoretical and practical significance of the present results, it is important to replicate the findings with an independent data set. Thus, we ran similar analyses with the pretest results from our recent hitherto unpublished WM training study. These analyses and their findings are reported in Experiment 2.

3. Experiment 2

The present experiment is a replication attempt of the first experiment. It follows closely the analysis plan of Experiment 1, but the data comes from a separate study with a new group of participants and a partly different WM test battery.

3.1. Participants and procedure

The data of Experiment 2 stems from the pretest of an unpublished randomized controlled online trial on working memory training (preregistration at https://osf.io/c9ygt). The study was approved by the Institutional Review Board of the Departments of Psychology and Logopedics, Åbo Akademi University, and it was conducted in accordance with the Helsinki Declaration. Following an extensive two-step prescreening procedure, 250 participants were invited to take part in the study through Prolific Academic (https://www.prolific.co/). Of those, 34 participants did not complete the pretest to the end, and an additional 15 participants reported cheating during task performance. This resulted in a final sample size of 201 participants that met our inclusion criteria, which were similar to those in Experiment 1 (i.e., English native speakers, no current psychiatric or neurological illnesses that affected the participant's daily life, no current use of central nervous system

(CNS) medication, and no current psychotropic drug use except tobacco, alcohol, and cannabis). The participants were 18–50-year-old with a mean age of 32.09 (SD = 8.27) and 56.5% (*n* = 114) of them were females. The mean education length was 16.13 years (SD = 3.35).

3.2. Materials

The seven WM tasks included in the pretest battery encompassed three n-back tasks (digits, letters, and colors), two Running memory tasks (letters, and colors), two simple span tasks (letters and colors), and two Selective Updating tasks (digits and colors).

3.2.1 N-back paradigm

We administered an n-back task with digits (NBD), letters (NBL) and colors (NBC). The items in each n-back variant were presented identically to those in Experiment 1, and the length was also the same.

3.2.2. Running memory

We administered two Running memory tasks, one with letters as stimuli (RML), and the other with colors (RMC). In both variants, stimulus presentation time was 1000 milliseconds with interstimulus interval at 500 milliseconds. The sequence lengths ranged from 4 to 11 items in both RML and RMC, and the item sequences were presented in a randomized order. The participants completed one trial of each sequence length. The length of these tasks was ca 5 minutes each.

3.2.3. Span

We administered two Forward simple span tasks, one with letters as stimuli (FSL), and the other with colors as stimuli (FSC). In both variants, participants are prompted with a sequence of items (presentation time 1000 milliseconds, interstimulus interval 500 ms), and the task is to recall the items in the order in which they were presented. The sequence lengths ranged from 4 to 9 in the FSL variant, and 4 to 10 in the FSC variant, and the item sequences were presented in a randomized order. The participants completed one trial of each sequence length. The length of these tasks was ca 5 minutes each.

3.2.4. Selective updating

The Selective updating (SU) paradigm builds upon the task introduced by Murty et al. (2011). In this study, we administered two tasks variants, one with digits (1 to 9) as stimuli (SUD) and one with colors (blue, yellow, red, green, purple, black, pink, orange, and gray) as stimuli (SUC). Besides materials, SUD and SUC were identical to each other (thus, the word 'item' pertains to both variants hereafter in the task description). In both tasks, five unrelated items were presented on the computer screen in a row of five boxes. The participants were instructed to memorize the item sequence. After this, the initial item sequence disappeared after which a new row of five boxes was displayed. Two of the new boxes contained new items, while three were empty. The participants were prompted to replace the old items with the items presented most recently in the memorized sequence, while maintaining the unchanged items in WM. In both SUD, and SUC, the participants completed 10 baseline trials (i.e., no updating stages), and 10 trials with three updating stages (i.e., replacement of old items with new ones). The participants were instructed to report the final item sequence. The order of the sequences was randomized, and the participants were unaware whether the next sequence would be a baseline sequence or an updating sequence. The initial item sequence was shown for 4000 milliseconds, followed by a 100-millisecond blank screen, after which the first updating stage was presented for 2000 milliseconds. The updating stage was once again followed by a 100-millisecond blank screen and the next updating stage. After all the updating stages had been presented (none in the baseline condition), a recall grid with horizontally aligned boxes containing the numbers from 1 to 9 (or the nine colors) appeared on-screen. The participants were to click on the numbers (or colors) in correct order (see [Fellman et al., 2018; Laine et al., 2018] for more technical details). The length of these tasks was ca 8 minutes each.

3.3. Dependent variables

As in Experiment 1, we first conducted CFAs on the whole dataset using summative data. For the n-back tasks, the dependent variables were the average level of n achieved across the 12 n-back blocks. For the Running memory and the Span tasks, the dependent variables were the total number of correctly recalled items in their correct position across all sequences. For the Selective updating tasks, the dependent variables were the total number of correctly recalled items in the updating trials.

Secondly, we split the sequences into three phases (initial, mid, final), roughly matched by length. The divisions were as follows: 4/4/4 in n-back (i.e., NBD, NBL, NBC); 2/2/3 in Running memory (i.e., RML and RMC), 2/2/2 in FSL, 2/2/3 in FSC, and 3/3/4 in the

Selective updating tasks (i.e., SUD and SUC). The dependent variable for the n-back tasks consisted of the highest n-back level achieved in the respective task phase. For the Span tasks, the dependent variables consisted of the proportion of correctly recalled items in correct serial order within the respective task phases (i.e., number of correctly recalled items/Total number of items). For the Running memory tasks and the Selective Updating tasks (including only the updating trials), the dependent variables were calculated by counting the correctly recalled items in correct position within the respective task phases.

3.4. Analysis plan in Experiment 2

We employed the same analysis plan and fit indices criteria as in Experiment 1 by comparing a g-factor model against a paradigm-specific model using the full dataset and a three-split analysis (see Fig. 5).

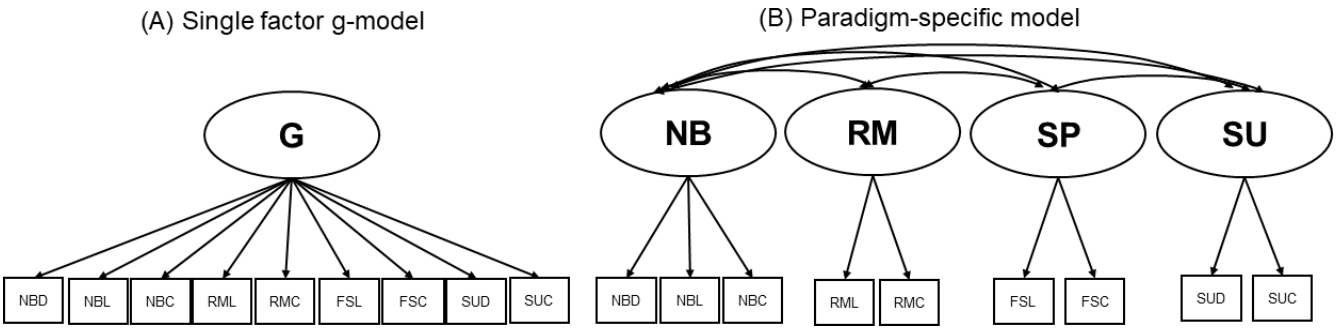

Figure 5. Candidate models for the analyses in Experiment 2. (A) A single g-factor model where all tasks load on a common factor; (B) A paradigm-specific four-factor model. Latent factors are shown in ovals and squares represent observed variables. G = g-factor; NB = N-back; RM = Running memory; SP = Span; SU = Selective updating; NBD = N-back with digits; NBL = N-back with letters; NBC = N-back with colors; RML = Running memory with letters; RMC = Running memory with colors; FSL = Forward span with letters; FSC = Backward span with colors; SUD = Selective updating of digits; SUC = Selective updating of colors.

3.5. Results of Experiment 2

The data were screened to identify univariate outliers (i.e., scores deviating >3.5 *SD*s from the sample mean on each task). Furthermore, color-blind participants performances in the color tasks were removed. These criteria resulted to 1.19% ($n_{observations}$ = 24) missing data in the full-data analysis, and 1.76%, and 1.17% ($n_{observations}$ = 70) of missing data in the three-split analysis. Descriptive statistics and pairwise correlations between tasks are summarized in Supplementary Materials, Appendix B.

3.5.1. Full-data analysis

The results from the CFAs on the full dataset (see Table 3) showed that the g-factor solution possessed a poor fit (χ^2 (27) = 254.906, CFI = 0.692 RMSEA = 0.205 (90% CI = 0.182, 0.228), SRMR = 0.105), whereas the paradigm-specific model showed a good fit (χ^2 (21) = 21.418, CFI = 0.997, RMSEA = 0.024 (90% CI = 0.000, 0.064), SRMR = 0.029). When comparing the g-factor model with the paradigm-specific model using the chi-square difference test, the paradigm-specific model provided a better fit to the data ($\Delta\chi^2$ = 176.99, Δdf = 6, p < 0.001) (see also Figure 6 depicting the factor loadings of the two models).

Table 3 Fit statistics for both models included in the confirmatory analyses on the full data in Experiment 2.

							Single-factor G vs. Paradigm-specific		
Model	χ^2	*df*	CFI	SRMR	RMSEA	AIC	$\Delta\chi^2$	Δdf	*p*
Single-factor G	254.906	27	0.692	0.105	0.205 [0.182, 0.228]	4571	176.992	6	< .001
Paradigm-specific	21.418	21	0.997	0.029	0.024 [0.000, 0.064]	4350			

Note. χ2 = Chi-square; *df* = degrees of freedom; Δ = difference; *p* = *p* value; CFI= comparative fit index; RMSEA = root mean square error of approximation (90% confidence intervals are given in square brackets); SRMR= standardized root-mean-square residual; AIC= Akaike information criterion.

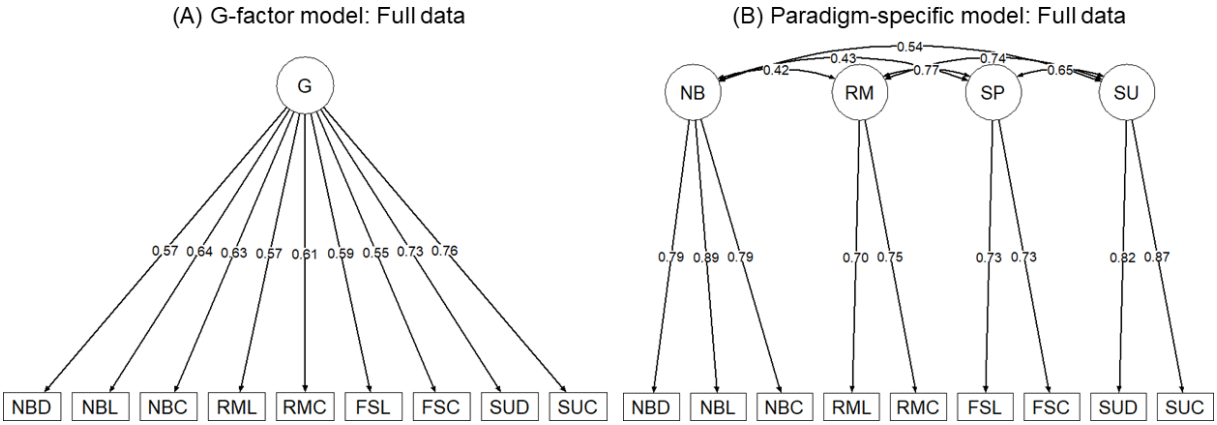

Figure 6. Full data: (A) A single g-factor model where all tasks load on a common factor; (B) A paradigm-specific four-factor model. Latent factors are shown in ovals and squares represent observed variables. G = g-factor; NB = N-back; RM = Running memory; SP = Span; SU = Selective updating; NBD = N-back with digits; NBL = N-back with letters; NBC = N-back with colors; RML= Running memory with letters; RMC= Running memory with colors; SUD = Selective updating with digits; SUC = Selective updating with colors.

3.5.2. Three-split analyses

The results from the CFAs on the **initial task phase** in the three-split analyses (see Table 4) showed that the g-factor solution had a poor fit (χ^2 (27) = 61.111, CFI = 0.838 RMSEA = 0.081 [90% CI = 0.055, 0.106], SRMR = 0.061), whereas the paradigm-specific model exhibited a good fit (χ^2 (21) = 25.09, CFI = 0.974, RMSEA = 0.037 [90% CI = 0.000, .072], SRMR = 0.040). When comparing the g-factor model with the paradigm-specific model, the paradigm-specific model provided a better fit to the data ($\Delta \chi^2$ = 33.82, Δ *df* = 6, *p* < 0.001) (see also Fig. 7A, and 7B depicting the factor loadings of the two models).

With respect to the **mid task phase**, the fit of the g-factor solution to the data was poor (χ^2 (27) = 118.846, CFI = 0.772, RMSEA = 0.086 (90% CI = 0.109, 0.155), SRMR = 0.086), whereas the paradigm-specific model revealed a good fit (χ^2 (21) = 19.253, CFI = 0.999, RMSEA = 0.007 (90% CI = 0.000, 0.057), SRMR = 0.029). When comparing the models using the chi-square difference test, the g-factor model was outperformed by the paradigm-specific model ($\Delta \chi^2$ = 99.621, Δ *df* = 6, *p* < 0.001) (see also Fig. 7C, and Fig. 7D depicting the factor loadings of the two models). However, in line with Experiment 1, the paradigm-specific model produced a covariance matrix that was not positive definite, stemming from a high correlation between the Running memory and Span factor (*r* = .96). This misspecification suggests that the exogenous variables belonging to these factors should be merged onto a common latent factor. However, for the sake of consistency, we refrained from computing additional models in Experiment 2 as well.

For the **final task phase**, the g-factor solution exhibited again a poor fit to the data (χ^2 (27) = 181.153, CFI = 0.744, RMSEA = 0.168 (90% CI = 0.145, 0.192), SRMR = 0.108), whereas the paradigm-specific model showed a good fit (χ^2 (21) = 18.711, CFI = 1.000, RMSEA = 0.000 (90% CI = 0.000, 0.056), SRMR = 0.025). When comparing the models using the chi-square difference test, the g-factor model was outperformed by the paradigm-specific model ($\Delta \chi^2$ = 112.694, Δ *df* = 6, *p* < 0.001) (see also Fig. 7E and Fig. 7F depicting the factor loadings of the two models).

Table 4 Fit statistics for both models included in the confirmatory analyses on the three-split data analyses in Experiment 2.

							Single-factor G vs. Paradigm-specific		
Model	χ^2	*df*	CFI	SRMR	RMSEA	AIC	$\Delta\chi^2$	Δdf	*p*
Initial task phase									
Single-factor G	61.111	27	0.838	0.061	0.081 [0.055, 0.106]	4919	33.828	6	< .001
Paradigm-specific	25.09	21	0.974	0.040	0.037 [0.000, 0.072]	4895			
Mid task phase									
Single-factor G	118.846	27	0.772	0.086	0.131 [0.109, 0.155]	4783	99.621	6	< .001
Paradigm-specific	19.253	21	0.999	0.029	0.007 [0.000, 0.057]	4695			
Final task phase									
Single-factor G	181.153	27	0.744	0.108	0.168 [0.145, 0.192]	4638	112.694	6	< .001
Paradigm-specific	18.711	21	1.000	0.025	0.000 [0.000, 0.056]	4488			

Note. $\chi 2$ = Chi-square; *df* = degrees of freedom; Δ = difference; *p* = *p* value; CFI= comparative fit index; RMSEA = root mean square error of approximation (90% confidence intervals are given in square brackets); SRMR= standardized root-mean-square residual; AIC= Akaike information criterion.

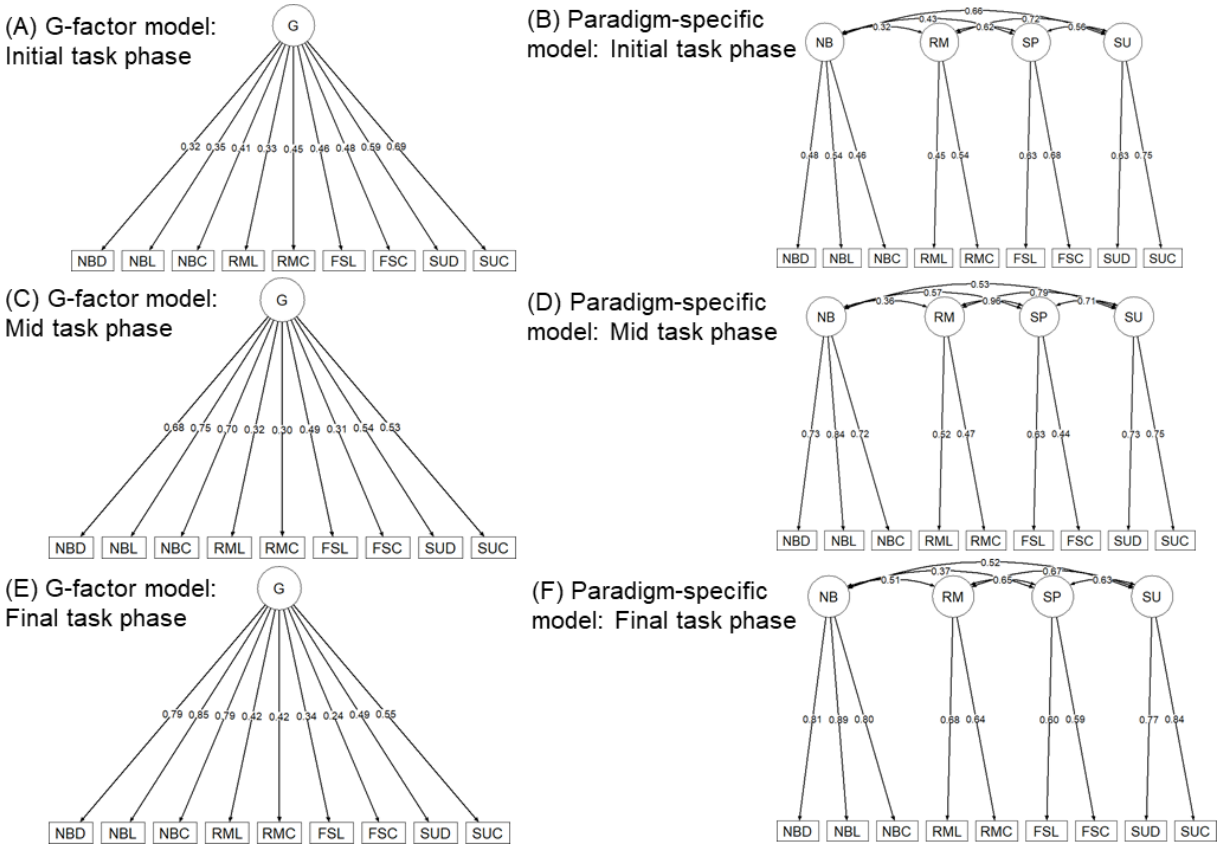

Figure 7. Three-split data: (A) A single g-factor model where all tasks load on a common factor on initial task performance; (B) A paradigm-specific model on initial task performance. (C) A single g-factor model where all tasks load on a common factor on mid task performance. (D) A paradigm-specific model on mid task performance (note that the high factor correlation between RM and SP produced a covariance matrix that was not positive definite) (E) A single g-factor model where all tasks load on a common factor on final task performance. (F) A paradigm-specific model on final task performance. Latent factors are shown in ovals and squares represent observed variables. G = g-factor; NB = N-back; RM = Running memory; SP = Span; SU = Selective updating; NBD = N-back with digits; NBL = N-back with letters; NBC = N-back with colors; RML= Running memory with letters; RMC= Running memory with colors; SUD = Selective updating with digits; SUC = Selective updating with colors.

3.5.3. Post hoc analyses

As in Experiment 1, there was a clear trend of the g-factor fit being initially better and then showing a decline, but the paradigm-specific model outperformed the g-factor model at all three stages. To examine the fit of the g-factor model at the very outset of task performance, post hoc analyses were conducted with the very first blocks in each task, except for the n-back task, where the first three blocks were used (see section 2.6.). At this pre-initial phase, the g-factor model showed good fit, (χ^2 (27) = 33.377, CFI = 0.925, RMSEA = 0.035 (90% CI = 0.000, 0.068), SRMR = 0.049), as did the paradigm-specific model (χ^2 (21) = 22.143, CFI = 0.971, RMSEA = 0.0275 (90% CI = 0.000, 0.065), SRMR = 0.043). However, the correlation between Running memory and Span in the paradigm-specific model was inadmissible (i.e., greater than 1), producing a covariance matrix that was not positive definite. The chi-square difference test showed that the two models did not differ in fit ($\Delta \chi^2$ (6) = 33.377, p = 0.154). The evolvement of model fit indexes over time is depicted in Figure 8.

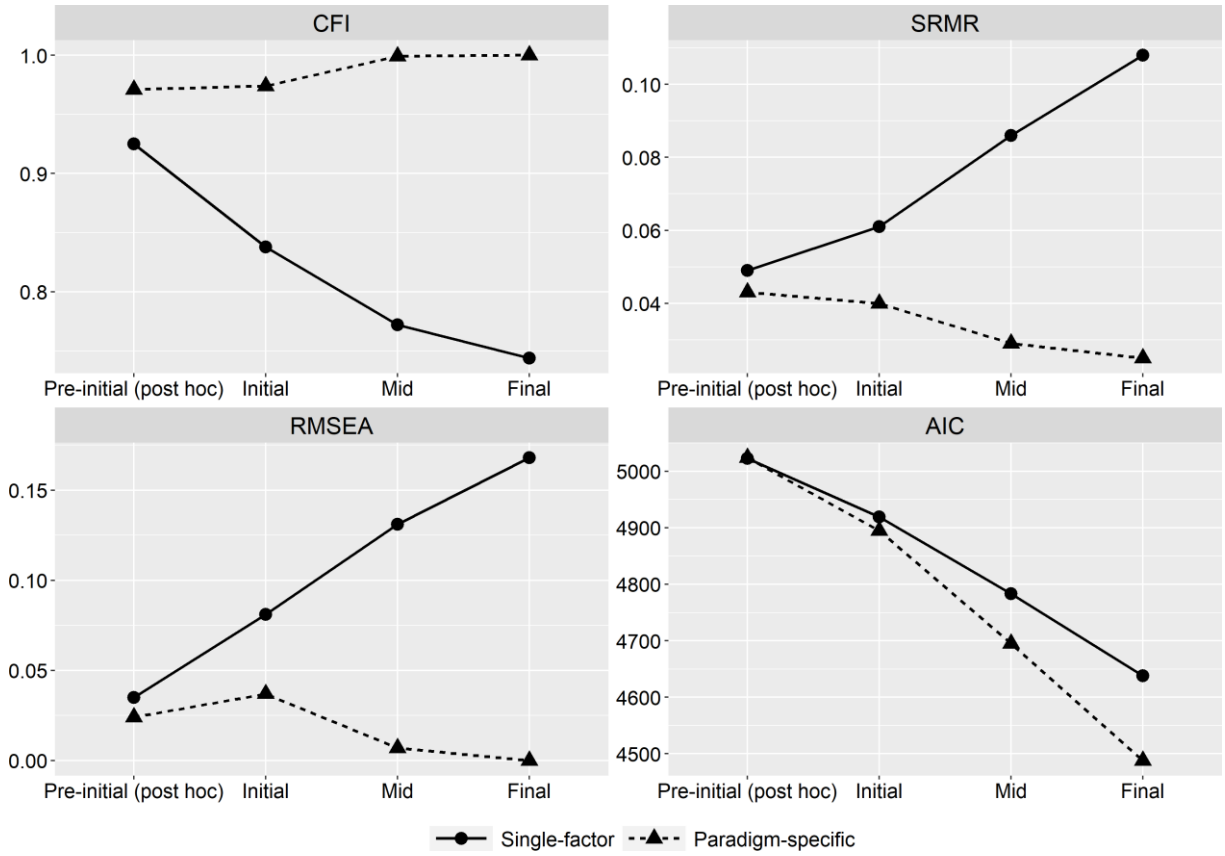

Figure 8. Model fit plot indices from pre-initial phase to final in Experiment 2. CFI= comparative fit index (larger is better); RMSEA = root mean square error of approximation (smaller is better); SRMR= standardized root-mean-square residual (smaller is better); AIC= Akaike information criterion (smaller is better).

3.6. Discussion (Experiment 2)

Experiment 2 successfully replicated the main results of Experiment 1 with new participants and with a slightly different set of WM tasks. As in the first experiment, the fit of the g-factor model declined steadily as the task progressed in the three-split analyses, whereas that of the paradigm-specific model largely stayed constant (see Figure 8). Contrary to our hypothesis, and similarly to Experiment 1, the paradigm-specific model outperformed the g-factor model throughout all three stages. A post-hoc analysis was again conducted to examine model fits at the very first phase of task performance. As in Experiment 1, at this earliest time point the fit indices of the paradigm-specific model were not significantly better than those of the g-factor model. We take this result and the subsequent decline in the g-factor model fit to support our adaptive cognition hypothesis that states that task-general executive resources are engaged particularly in the beginning of WM tasks.

4. General discussion

The present study addressed a fundamental untested assumption in cognitive testing, namely that the cognitive processes underlying complex task performance remain the same throughout the whole task. In contrast, cognitive skill learning literature suggests that performing a task is an adaptive learning process that proceeds from task-general metacognitive and control processes to more automatic and task-specific ones (Chein & Schneider, 2012; Schneider & Chein, 2003; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). Here we applied this approach for the first time in cognitive WM tasks, the administration of which takes only a few minutes per task. During these brief task periods, we expected a shift in the latent structure of WM task performances as the testing period unfolded. More specifically, we hypothesized that during the early stages, different WM task performances would show highest covariance, reflecting the engagement of task-general metacognitive and executive resources, such as generating and implementing strategies (e.g., Fellman et al., 2020; Laine et al., 2018). This covariance would be seen in the fit indices of a g-factor model that would be highest in the beginning of task performance and then show a downward slope towards the end when the task is becoming more familiar. At the same time, we expected to see an increase in the fit of the paradigm-specific model, reflecting the automatization of the tasks. In contrast to this adaptive cognition hypothesis, the static cognition hypothesis, implicitly applied in cognitive testing, assumes that the cognitive architecture underlying complex cognitive task performance remains stable.

The results from the two independent data sets analyzed in Experiments 1 and 2 supported the adaptive cognition hypothesis in that the fit of the g-factor model was best at the outset and steadily declined as the task progressed. This indicates that task-general executive processes are engaged especially at the initial stages of task performance, and less so at the latter stages. This phenomenon is quite short-lived in WM tasks, as the fit of the g-factor model was comparable to that of the nested paradigm-specific model only at the very earliest pre-initial task phase (see Figures 4 and 8 and the post hoc analyses), while the paradigm-specific model was better than the g-factor model at all consecutive task stages. As predicted by the adaptive cognition hypothesis, the latent factor structure changed as the task progressed. This change was mainly due to the decline in the fit of the g-factor model; the fit of the paradigm-specific model was acceptable throughout task performance. Hence, this did not confirm our expectation that the fit of the paradigm-specific model would increase as paradigm-specific skills are learnt.

A possible reason why the paradigm-specific model had a good fit already in the beginning of task performance is that the present WM tasks were not problem-solving tasks but instead straightforward tests with clear performance instructions coupled with practice trials. Thus, the task-specific processes can be formed quickly, although their application requires task-general cognitive control in the very beginning. However, the most active control phase is short-lived, as suggested by the rapid decline in the fit of the g-factor model. The decline of task-general control would in turn reflect the buildup of task-specific skill (Chein & Schneider, 2012). So, how can automatization increase while the fit of the paradigm-specific model is constant? This could be explained in terms of the skill learning model of Taatgen (2013), where automatization is represented as the chaining of primitive processing elements, called PRIMs. A single PRIM could consist of a rule or algorithm such as "check for visual input" or "copy visual input to retrieval". In this model, individual PRIMs become conjoined through practice into a single unit that can be automatically performed. For example, suppose that a task requires mental operations A, B, C, and D. At the outset, these operation steps are effortfully performed one after the other, but during repeated task performance, they start to form a "production rule" in long-term memory (A-B-C-D), which can be automatically activated. In the production rule, the mental operations are identical to what they were before automatization and what changes is merely how the operations are connected. This could explain why the fit of the paradigm-specific model in the present experiments stayed constant: the operations required by the tasks remain constant, and what changes is how they are linked in long-term memory. This change would not be reflected as a change in the fit of the paradigm-specific model, but instead only as a change in effortful, task-general processing. In short, the task-specific mechanisms in the present experiments were apparently formed very quickly, and what changed was how effortfully they were recruited.

The main finding in both experiments was that the role of task-general processes diminished as the task progressed. Thus, the underlying cognitive architecture of complex WM task performance changed from task-general to task-specific during the few minutes each task took. Following the skill learning approach, we assume that this reflects the rapidly diminishing role of metacognitive and control processes, followed by a gradual buildup of task-specific automatic skill. This contrasts with the traditional assumption that cognitive tasks tap on static cognitive constructs. The results have important implications, but before discussing them, it is worth noting some limitations of the current study.

4.1. Limitations

An important limitation is that our analytical method (CFA) only yields indirect evidence of the processes underlying task performance. Whereas it is clear that the g-factor model reflects task-general processes and the paradigm-specific model task-specific processes, the results leave open to what extent the former are effortful and the latter automatic; this is merely hypothesized on the basis of the skill learning theory. We assume that the g-factor model reflects metacognitive and controlled processes, whereas the paradigm-specific model (at least in the later stages where the g-factor model had a poor fit) reflects more automatic, lower-level processes. Whether this is in fact the case would require more direct measures of automatization. One possibility could be to follow up the speed of the basic stimulus-response mapping required by a task. To take a concrete example, in an adaptive n-back task one could employ intermittent 1-back blocks that have a minimal WM load and track the speed-up of reaction times on these blocks. This could tap on the emergence of task-specific automatic production rules at the sensory-motor level (Taatgen, 2013). Another possible option would be the use of some simple intermittent secondary task. Due to the task-initial high executive load on the complex and unfamiliar primary task, the secondary task should create clear interference early on. However, this interference effect should be attenuated over time, reflecting increasing automaticity of the primary task that frees executive resources for the secondary task.

The interpretation of the present results as reflecting skill learning is supported by independent evidence from our two previous microgenetic studies that focus on strategy use in memory tasks (Waris et al., 2020, 2021). These results showed that changes in strategies take place mainly during the first 2-3 blocks of a task, which indicate short-lived metacognitive and control processing. After the first blocks, strategies became more stabilized while performance was steadily enhanced. This could reflect increased automatization in the employment of the strategies.

Another limitation relates to probable inter-task and inter-individual differences in when the peak executive load occurs. To take an example of the former, in simple span tasks executive load is most likely rather weak and arguably peaks at an early stage, whereas the n-back task is more complex and assumedly loads on the executive system for a longer period. The length and intensity of these executively taxing task-initial periods may also vary considerably between participants, depending on their individual characteristics. The current analytical method with rough splits of the task periods at a group level does not take these factors into account. They could be better analyzed by using multilevel methods where

individual slopes for tasks and participants can be specified, but these in turn cannot presently yield latent constructs. A related limitation is that the time resolution in our analysis is very coarse compared to proper time-series analyses. This limitation is due to the use of CFA, which is based on individual differences and requires large datasets. Again, the advantage of CFA is that its results are comparable to previous studies with the same method that have targeted the cognitive architecture underlying WM performance.

A final limitation of the present study is that the tasks stemmed from previous cognitive training studies and may not be ideal for the present type of analysis. For example, the present n-back tasks were adaptive which lessens the degree to which they can be automatized, we had practice rounds that most probably "ate up" part of the initial executive engagement peak, and span length variation between trials in the span tasks added variability to the results. Moreover, the length of the tasks varied substantially with the n-back task being the longest due to its adaptivity. These factors should be considered in future studies. The fact that the main findings were nevertheless replicated in two independent datasets with different sets of tasks speaks for the robustness of the results.

4.2. Theoretical and practical implications

Cognitive task performance is an adaptive process. The present results indicate that the skill learning approach is relevant also to psychometric tasks, indicating that even during the short time spans of these tasks, processing shifts from task-general control processes to automatic skills. The results are supported by independent evidence from the recent microgenetic studies by Waris et al. (2020, 2021), which likewise indicate a very early engagement of metacognition and cognitive control in the form of strategy development in memory tasks. This overlaps in time with the rapid diminishment of the g-factor model fit we observed in the present two experiments. We expect the shift from task-general to task-specific processes to apply to any complex cognitive task, not only WM tasks. Of particular interest here are executive functions (EF). In their highly influential study, Miyake and colleagues (2000) divided EF into three latent factors, namely inhibition, shifting, and WM updating/monitoring. Future research should run the present type of analyses with the kinds of tasks Miyake et al. (op. cit.) used. Whereas Miyake et al. employed summative scores and found that the EF system shows both unity and diversity (in that there are separable but correlated factors), we expect that the inclusion of the temporal dimension would reveal a shift from an initially higher unity towards diversity as the tasks progress.

Psychometrics should take the temporal dimension of task performance into account. Our results indicate that the task-general processes are most prominent in the beginning of task performance, assumedly reflecting higher engagement of executive and metacognitive processes at that point. This phase of high executive load may pass quickly. Thus, if the aim is to measure EF, the researcher should focus on the first stages of task performance; the later stages may reflect task-specific and partly automatic processes that no longer engage the executive system. More generally, cognitive testing should take time-structured intraindividual variability better into account, as suggested recently by Gonthier and Roulin (2020). One can speculate whether within-task learning curves and their slopes might be more sensitive measures of individual EF than the currently used summative scores. More advanced methods to address intraindividual diachronic changes include dynamic systems modeling or statistical methods such as generalized additive mixed models (GAMM; Gonthier & Roulin, 2020).

Different stages of task performance may show variable convergent and ecological validity. We showed that the g-factor model was most prominent in the beginning of the task. Under the assumption that this provides a purer measure of EF engagement, one could ask whether task-initial performance also correlates more strongly with other goal-directed adaptive behaviors than the commonly used summative scores that also encompass later phases of task performance. For instance, the initial stages of WM processing might correlate with the initial stages of other EF tasks, or with general fluid intelligence measures. The initial stages of complex cognitive task performances could also show stronger correlations with measures of everyday behaviors such as academic achievement. If this holds, it would have important implications for the application of cognitive tests.

Focus in cognitive testing should be moved from static constructs to adaptive processes. From the perspective of WM research particularly, the present results converge with literature that questions the notion of WM capacity. We agree with Simmering and Perone (2013) who argue that "WM capacity" is not capacity in the traditional meaning of the term (e.g., number of slots in a storage system), but the final product of multiple cognitive systems that operate during task performance. Thus, WM capacity is not a constant individual feature across tasks, and storage cannot be separated from other processes. Whereas Simmering and Perone (op. cit.) focus on the dynamicity of WM from a developmental perspective, our results indicate that change occurs also within a single testing session. More generally, the present line of thinking suggests that there are no static cognitive constructs measurable by specific tasks, but instead cognitive processes evolve in interaction with the

task demands. The focus of classical cognitive psychology on static constructs could be a reason why mapping cognitive functions to brain processes has largely been unsuccessful (cf. Poldrack, 2010; Poldrack & Yarkoni, 2016). If the cognitive processes employed by a task change, the corresponding neural processes must change as well (cf. Badre et al., 2010). Thus, instead of mapping static cognitive constructs to brain mechanisms, a more viable goal might be to map temporally unfolding cognitive processes to likewise unfolding neural processes, both of which interact with the organism's environment (see, e.g., Bassett & Sporns, 2017; Bassett et al., 2011; Braun et al., 2015).

Behavioral testing partly creates what it measures. Finally, in a broader and more philosophical perspective, the results support the general "Spencerian" hypothesis that cognition is a process that adapts to the context. This implies that cognitive mechanisms may not exist "out there" at the latent level, waiting to be measured. Instead, cognitive testing partly creates what is being measured, as the cognitive system adapts to the test. This is emphasized by theories of situated or embodied cognition (e.g., Anderson, 2014). To illustrate by analogy, we may assume that the brain carries no general schema for loading the dishwasher prior to being exposed to that activity; there is only a general capacity to learn perceptual-motor processes, individual instances of which are near-infinite. Is there a mechanism in the brain for conducting a WM task like n-back prior to being exposed to that task? If not, then what does exist prior to learning the task? It is commonly assumed that at least WM exists independently of our measuring it, but it can only be assessed through cognitive tasks, all of which elicit learning. The present results suggest that WM has a task-general component, but how it is manifested in behavior depends on the interaction between the person and the environmental demands.

4.3. Conclusions

The present results challenge the assumption that the human cognitive architecture remains static during the brief time periods (at the level of minutes) that most cognitive tasks take. Using WM tasks, we revealed a shift from task-general to task-specific processing during the short task periods. This can be conceptualized along the lines of the skill learning theory as a shift from global metacognitive and executive processes to task-specific and more automatic processes. This implies that what cognitive tasks measure changes over time and motivates a stronger focus on within-task intraindividual variability and learning, instead of the mere use of static summative measures in cognitive testing. More generally, the results support the view that cognition is an adaptive process that, to paraphrase Spencer, molds into

the shape of the environment. It follows that it is impossible to behaviorally measure a cognitive process without affecting it, as the system quickly adapts to the context through learning task-specific skills. A detailed analysis of cognitive task performance as it unfolds over time provides a rich but underutilized source of data that promises to further our understanding on what cognitive tasks measure.

References

Anderson, M. L. (2014). *After Phrenology. Neural Reuse and the Interactive Brain*. Cambridge: MIT Press.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin and Review*, *22*(2), 366–377. https://doi.org/10.3758/s13423-014-0699-x

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, *66*(2), 315–326. https://doi.org/10.1016/j.neuron.2010.03.025

Bassett, D. S., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, *20*(3), 353–364. https://doi.org/10.1038/nn.4502

Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., & Grafton, S. T. (2011). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(18), 7641–7646. https://doi.org/10.1073/pnas.1018985108

Bledowski, C., Rahm, B., & Rowe, J. B. (2009). What "works" in working memory? Separate systems for selection and updating of critical information. *Journal of Neuroscience*, *29*(43), 13735–13741. https://doi.org/10.1523/JNEUROSCI.2547-09.2009

Braun, U., Schäfer, A., Walter, H., Erk, S., Romanczuk-Seiferth, N., Haddad, L., … Bassett, D. S. (2015). Dynamic reconfiguration of frontal brain networks during executive cognition in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(37), 11678–11683. https://doi.org/10.1073/pnas.1422487112

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin*, *107*(2), 260–273. https://doi.org/10.1037/0033-2909.107.2.260

Chein, J. M., & Schneider, W. (2012). The brain's learning and control architecture. *Current Directions in Psychological Science*, *21*(2), 78–84. https://doi.org/10.1177/0963721411434977

Colom, R., Shih, P. C., Flores-Mendoza, C., & Quiroga, M. Á. (2006). The real relationship between short-term memory and working memory. *Memory*, *14*(7), 804–813. https://doi.org/10.1080/09658210600680020

Fellman, D., Soveri, A., Viktorsson, C., Haga, S., Nylund, J., Johansson, S., … Laine, M.

(2018). Selective updating of sentences: Introducing a new measure of verbal working memory. *Applied Psycholinguistics*, *39*(2), 275–301. https://doi.org/10.1017/S0142716417000182

Fellman, Jylkkä, Waris, Soveri, Ritakallio, Haga, … Laine. (2020). The role of strategy use in working memory training outcomes. *Journal of Memory and Language*, *110*(104064).

Forsberg, A., Fellman, D., Laine, M., Johnson, W., & Logie, R. H. (2020). Strategy mediation in working memory training in younger and older adults. *Quarterly Journal of Experimental Psychology*, *73*(8), 1206–1226. https://doi.org/10.1177/1747021820915107

Gathercole, S. E., Dunning, D. L., Holmes, J., & Norris, D. (2019). Working memory training involves learning new skills. *Journal of Memory and Language*, *105*, 19–42. https://doi.org/10.1016/J.JML.2018.10.003

Gonthier, C., & Roulin, J. L. (2019). Intraindividual strategy shifts in Raven's Matrices, and their dependence on working memory capacity and need for cognition. *Journal of Experimental Psychology: General*, *149*(3), 564–579. https://doi.org/10.1037/xge0000660

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60. https://doi.org/10.21427/D79B73

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Karbach, J., & Verhaeghen, P. (2014). Making working memory work: A meta-analysis of executive-control and working memory training in older adults. *Psychological Science*, *25*(11), 2027–2037. https://doi.org/10.1177/0956797614548725

Kassai, R., Futo, J., Demetrovics, Z., & Takacs, Z. K. (2019). A meta-analysis of the experimental evidence on the near- and far-transfer effects among children's executive function skills. *Psychological Bulletin*, *145*(2), 165–188. https://doi.org/10.1037/bul0000180

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358. https://doi.org/10.1037/h0043688

Laine, M., Fellman, D., Waris, O., & Nyman, T. J. (2018). The early effects of external and internal strategies on working memory updating training. *Scientific Reports*, *8*(1), 1–12.

https://doi.org/10.1038/s41598-018-22396-5

Mackintosh, N. J., & Bennett, E. S. (2003). The fractionation of working memory maps onto different components of intelligence. *Intelligence, 31*(6), 519–531. https://doi.org/10.1016/S0160-2896(03)00052-7

Melby-Lervåg, M, Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer": Evidence from a meta-analytic review. *Perspectives on Psychological Science, 11*(4), 512–534. https://doi.org/10.1177/1745691616635612

Melby-Lervåg, Monica, & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49*(2), 270–291. https://doi.org/10.1037/a0028228

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

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. https://doi.org/10.1006/cogp.1999.0734

Murty, V. P., Sambataro, F., Radulescu, E., Altamura, M., Iudicello, J., Zoltick, B., … Mattay, V. S. (2011). Selective updating of working memory content modulates meso-cortico-striatal activity. *NeuroImage, 57*(3), 1264–1272. https://doi.org/10.1016/j.neuroimage.2011.05.006

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and Self-Regulation* (pp. 1–14). Boston: Springer.

Poldrack, R. A. (2010). Mapping mental function to brain structure: How can cognitive neuroimaging succeed? *Perspectives on Psychological Science, 5*(6), 753–761. https://doi.org/10.1177/1745691610388777

Poldrack, R. A., & Yarkoni, T. (2016). From brain maps to cognitive ontologies: Informatics and the search for mental structure. *Annual Review of Psychology, 67*, 587–612. https://doi.org/10.1146/annurev-psych-122414-033729

R Core Team (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. Available online at https://www.R-project.org/.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(1), 1–36. https://doi.org/10.18637/jss.v048.i02

Sala, G., & Gobet, F. (2017). Working memory training in typically developing children: A meta-analysis of the available evidence. *Developmental Psychology*, *53*(4), 671–685. https://doi.org/10.1037/dev0000265

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: behavior, theory, and biological mechanisms. *Cognitive Science*, *27*(3), 525–559. https://doi.org/10.1207/s15516709cog2703_8

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist*, *50*(2), 138–166. https://doi.org/10.1080/00461520.2015.1036274

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, *125*(1), 4–27. https://doi.org/10.1037//0096-3445.125.1.4

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Simmering, V. R., & Perone, S. (2013). Working memory capacity as a dynamic process. *Frontiers in Psychology*, *3*(567). https://doi.org/10.3389/fpsyg.2012.00567

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-016-1217-0

Spencer, H. (1855). *Principles of Psychology*. London: Williams and Norgate.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, *120*(3), 439–471. https://doi.org/10.1037/a0033138

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. https://doi.org/10.1016/j.cogpsych.2014.01.003

Vuong, L. C., & Martin, R. C. (2014). Domain-specific executive control and the revision of misinterpretations in sentence comprehension. *Language, Cognition and Neuroscience*,

29(3), 312–325. https://doi.org/10.1080/01690965.2013.836231

Waris, O., Fellman, D., Jylkkä, J., & Laine, M. (2020). Stimulus novelty, task demands, and strategy use in episodic memory. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/1747021820980301

Waris, O., Jylkkä, J., Fellman, D., & Laine, M. (2021). Spontaneous strategy use during a working memory updating task. *Acta Psychologica*, *212*, 103211. https://doi.org/10.1016/j.actpsy.2020.103211

Waris, O., Soveri, A., Ahti, M., Hoffing, R. C., Ventus, D., Jaeggi, S. M., … Laine, M. (2017). A latent factor analysis of working memory measures using large-scale data. *Frontiers in Psychology*, *8*(1062). https://doi.org/10.3389/fpsyg.2017.01062

Weicker, J., Villringer, A., & Thöne-Otto, A. (2016). Can impaired working memory functioning be improved by training? A meta-analysis with a special focus on brain injured patients. *Neuropsychology*, *30*(2), 190–212. https://doi.org/10.1037/neu0000227

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*(433). https://doi.org/10.3389/fpsyg.2013.00433