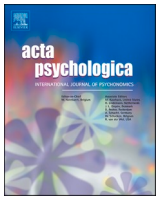
Acta Psychologica 212 (2021) 103211

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



Spontaneous strategy use during a working memory updating task

Otto Waris [a,b,c,*], Jussi Jylkkä [a], Daniel Fellman [a,d], Matti Laine [a,e]

[a] Department of Psychology, Åbo Akademi University, Turku, Finland
[b] Department of Child Psychiatry, Research Centre for Child Psychiatry, University of Turku, Turku, Finland
[c] INVEST Research Flagship, University of Turku, Turku, Finland
[d] Department of Applied Educational Science, Umeå University, Umeå, Sweden
[e] Turku Brain and Mind Center, University of Turku, Turku, Finland

ARTICLE INFO

Keywords:
Working memory
N-back
Strategy
Skill learning
Routine

ABSTRACT

Cognitive skill learning postulates strategy generation and implementation when people learn to perform new tasks. Here we followed self-reported strategy use and objective performance in a working memory (WM) updating task to reveal strategy development that should take place when faced with this novel task. In two pre-registered online experiments with healthy adults, we examined short-term strategy acquisition in a ca 20–30-minute adaptive n-back WM task with 15 task blocks by collecting participants' strategy reports after each block. Experiment 1 showed that (a) about half of the participants reported using a strategy already during the very first task block, (b) changes in selected strategy were most common during the initial task blocks, and (c) more elaborated strategy descriptions predicted better task performance. Experiment 2 mostly replicated these findings, and it additionally showed that compared to open-ended questions, the use of repeated list-based strategy queries influenced subsequent strategy use and task performance, and also indicated higher rates of strategy implementation and strategy change during the task. Strategy use was also a significant predictor of n-back performance, albeit some of the variance it explained was shared with verbal productivity that was measured with a picture description task. The present results concur with the cognitive skill learning perspective and highlight the dynamics of carrying out a demanding cognitive task.

1. Introduction

Cognitive functions are typically perceived to be relatively stable, latent constructs that can be assessed through performance-based cognitive tasks (e.g., Baddeley, 2010; Friedman & Miyake, 2017; Miyake et al., 2000). In contrast, the skill learning approach to cognition postulates that the cognitive system is essentially adaptive, aiming to optimize performance in a given task (Schneider & Chein, 2003; Chein & Schneider, 2012; Taatgen, 2013; see also Hasson et al., 2020). Thus, from a skill learning perspective, even a single test session with a cognitive task is already an adaptive process, where the participant must learn how to perform that task. Chein and Schneider (2012) suggested that skill learning takes place in three phases. Upon encountering a novel task, the cognitive system enters the Formation stage, where the metacognitive system establishes strategies and behavioral routines that enable task performance. These processes are effortfully put into practice in the Controlled Execution phase, which relies on the cognitive control system. Finally, task performance starts to gradually become more automatic and modular in the Automatic Execution phase, whereby the resources of the Metacognitive and Cognitive Control systems are freed to other tasks (for similar accounts of learning, see Baars, 1988, 2002; Dehaene & Changeux, 2011). From this perspective, cognitive task performance is not monolithic but reveals a dynamic chain of events if analyzed in sufficient detail.

In accordance with the skill learning theory, previous research has indicated that performance in many cognitive tasks improves over repeated test sessions, and even within a single test session (e.g., Calamia et al., 2012; Goldberg et al., 2015; Soveri et al., 2018). To better understand what a task measures, it is important to know what processes underlie such learning effects. Hence, in the present study, we focused on spontaneous strategy use in a widely utilized working memory (WM) updating task, namely the n-back task (Kirchner, 1958). The formation and implementation of strategies is central in Chein and Schneider's (2012) first two stages of skill learning, namely the Formation and Controlled Execution phases. Strategies can be defined as conscious and effortfully created conceptual rules that can modulate lower-level

* Corresponding author at: Department of Psychology, Åbo Akademi University, Biskopsgatan 3, 20500 Turku, Finland.
E-mail address: owaris@abo.fi (O. Waris).

https://doi.org/10.1016/j.actpsy.2020.103211
Received 17 June 2020; Received in revised form 8 September 2020; Accepted 23 October 2020
Available online 18 November 2020

processes. An individual's consciously chosen method for performing a task can rely on cognitive mechanisms responsible for problem-solving and utilize information in episodic memory, such as previous experiences of mnemonics.

The widely used n-back task paradigm that we employed in the present study prompts participants to decide whether the current element (i.e., stimulus item) matches the one presented n trials ago, thus requiring continuous updating of incoming stimuli in WM. There were several reasons for choosing the n-back task for closer scrutiny: (a) it is a commonly used WM measure, (b) it is well suited for strategy analysis, as it is a novel task that calls for the generation and implementation of strategies or routines (Gathercole et al., 2019), and (c) it can serve as one model for the hitherto largely unexplored strategy evolvement during WM task performance. The last point is also of wider interest, as WM task practice in training paradigms has recently been argued to represent cognitive skill learning (Fellman et al., 2020; Gathercole et al., 2019; Laine et al., 2018). As noted above, cognitive skill learning is thought to progress in stages, where initial strategy generation and implementation is followed by a gradual development of task routine (Chein & Schneider, 2012; Taatgen, 2013). Cognitive skill learning models do not specify the timing of these stages for any specific task, but one could assume that in a task such as adaptive n-back that does not require complex problem solving or very advanced skills to start with, strategy implementation could be quite rapid, followed by a more stable use of the strategy one has settled with as the task becomes more familiar. Thus, we expected that strategy acquisition and stabilization could be observable even within a single n-back test session.

Previous studies collecting participant reports have indicated that strategy use such as active rehearsal, stimulus grouping, or use of associations, is quite common in WM tasks. Spontaneous strategy use and its positive associations with WM performance have been documented in both simple and complex span tasks (Bailey et al., 2009, 2008, 2011; Dunlosky & Kane, 2007; Engle et al., 1990; Friedman & Miyake, 2004; Kaakinen & Hyönä, 2007; McNamara & Scott, 2001; Morrison et al., 2016) as well as in n-back (Fellman et al., 2020; Forsberg et al., 2020; Laine et al., 2018). The directionality of the relationship between strategy use and WM performance cannot be established by examining self-generated strategy use, but studies that have manipulated effective strategy use through external instructions speak for a causal relationship between strategy use and WM performance (Bailey et al., 2014; Borella et al., 2017; Carretti et al., 2007; Fellman et al., 2020; Forsberg et al., 2020; Laine et al., 2018). While these studies show that strategy use is an important aspect of WM performance that can in part explain the large inter-individual differences in WM measures, they have not examined strategy use block-by-block within a single WM test session, but rather reported aggregated strategy results. It is this very early phase of skill learning, assumed to unfold across task blocks right from the start of the task, that we aimed to reveal in two pre-registered online experiments. For this purpose, we recorded strategy self-reports and objective task performance after each task block in a 15-block adaptive single n-back task with digits, a widely used WM updating measure.

Our dual experiment study had several aims, all of them being related to strategy use during WM task performance. First, we sought to elucidate how fast participants report strategy use in the n-back task and how strictly they stick to a specific strategy from one task block to another. Second, to verify the relevance of self-reported strategy use for success on the task, we examined the relationship between strategy use (type and level of detail of a self-generated strategy) and objective task performance. These issues were examined in Experiment 1 and then subjected to replication in Experiment 2. Third, by employing a between-subjects design in Experiment 2, we examined a methodological issue, namely the effect of strategy inquiry format (open-ended question vs. list-based strategy query that provides a description of major strategies) on strategy use and objective WM task performance. Fourth, we probed the predictive value of selected variables for objective WM performance. Metacognitive ability as measured by the Metacognitive Awareness Inventory (Harrison & Vallin, 2017; Schraw & Dennison, 1994) was chosen as a predictor in Experiment 1 because the strategy generation phase in skill learning has been related to metacognition (e.g., Chein & Schneider, 2012). In Experiment 2, we introduced three other predictors. A written picture description task was used as a proxy for verbal productivity. We surmised that it could in part account for strategy employment that is likely to be strongly guided by the language system. Self-reported employment of internal memory aids in everyday life (as assessed by the Memory Aids Questionnaire, Chouliara & Lincoln, 2015) could also be associated with strategy use and higher performance on our experimental WM task. Finally, development of task routine was queried with a Likert-scale assessment on the ease of responding in the n-back task. As noted above, cognitive skill learning also entails gradual development of routine that would in part explain progress on the task.

2. Experiment 1

In this experiment, we investigated how fast participants develop strategies in the n-back task and how quickly strategy use stabilizes. We also analyzed the relationships between strategy use (self-generated strategy type and level of detail in the strategy description) and objective task performance, and assessed whether metacognitive ability predicted strategy use and n-back performance. The pre-registration of this experiment can be found at http://aspredicted.org/blind.php?x=fb8ek8.

2.1. Methods

2.1.1. Ethics statement

The experiment was approved by the joint ethics committee of the Departments of Psychology and Logopedics at Åbo Akademi University. We obtained informed consent from all participants, participation was anonymous, and we informed participants of their right to withdraw from the experiment at any time.

2.1.2. Procedure

This fully online experiment consisted of a background questionnaire, an adaptive digit n-back task with open-ended strategy queries after each task block, and a posttest questionnaire. The background questionnaire probed demographics like age and gender as well as certain exclusion criteria such as the presence of neurological illnesses (see section Participants, below). It also contained the self-rated 19-item shortened version of the Metacognitive Awareness Inventory (MAI; Schraw & Dennison, 1994; Harrison & Vallin, 2017). The MAI produces two interlinked factors: knowledge of cognition that taps respondents' knowledge and awareness of their thought processes, and regulation of cognition that taps control over thought processes through, for example, planning and monitoring (Brown & Palincsar, 1982). The posttest questionnaire contained questions about participants' n-back performance (e.g., an n-back strategy questionnaire, use of external tools, motivation, and effort). The n-back task with the block-wise strategy reports took approximately 20–30 min, and the whole experiment including the questionnaires took about 30–40 min.

2.1.3. Participants

We used the crowdworking site Prolific (https://www.prolific.ac/) to recruit 18–50 year-old participants from the United Kingdom or the USA with the help of Prolific's built-in prescreening tool. Participants were paid 3.33£ for taking the study. One hundred and ninety-nine participants completed the whole experiment. Sixty-eight of them were excluded for the following reasons: neurological illness ($n = 9$), psychiatric illness ($n = 35$), neurodevelopmental disorder ($n = 5$), medication or drugs that affect the CNS (excluding tobacco, alcohol, and cannabis products) ($n = 5$), consuming >9 units of alcohol on the previous day ($n = 3$), never reaching the 2-back level in the n-back task ($n =$

2), reporting previous experience of the n-back ($n = 6$), and reporting the use of external help (note-taking) to solve the n-back task ($n = 3$).[1] Finally, one additional participant was excluded as the user account had later been banned by Prolific for multiple account ownership. This gave us a final sample of 130 participants (see Table 1 for descriptives).

2.1.4. The adaptive n-back task

The adaptive digit n-back task in Experiment 1 required participants to recall whether the currently shown single digit (1-9) matched the digit presented n digits ago. If the digits matched, they were to press the n-key on the computer keyboard, and if they did not match, the m-key was to be pressed. For example, in the 3-back sequence 4-9-2-6-9-3, the first three digits (4-9-2) cannot be matched to anything, the digit 6 does not match the digit presented 3 steps back (which was a 4), the second 9 is a match, and the 3 is a no-match. The stimuli appeared on-screen one at a time at the center of the web browser window. Every digit was visible for 1500 ms and the digits were separated by a fixation cross that was shown for 450 ms. Each response was to be given within 1950 ms (stimulus exposure + fixation time). The task comprised 15 blocks of 20 + n items. Each block consisted of 6 target, 10 no-target, and 4 lure items. Target items matched the item presented n items ago, while no-target items and lures did not match the item shown n items ago. The difference between no-targets and lures was that lures matched the item presented n+1 or n-1 items back.[2] Lures were included to hamper possible familiarity-based responding (see Szmalec et al., 2011).

Each block began with the task instructions, followed by 20 + n stimulus items. After each block, the participants were asked to describe in their own words and with as much detail as possible any strategy that they had used during that block. After typing in a response, the result screen was displayed. It contained the number of correct responses (e.g., 18/20), a short verbal comment on the participant's performance, and a mention of the level of n in the next block. If 18 or more of the responses were correct, the level of n increased by one in the next block (up to a maximum of 15-back). If 15-17 responses were correct, the level of n remained the same; and if less than 15 responses were correct, the level of n was decreased by one (1-back being the minimum). For the purpose of this study, only accuracy rates were evaluated. The average level of n achieved in the 15 n-back blocks was used as the outcome variable. Average level of n was chosen over maximum level achieved due to its wider distribution.

After completing the whole n-back task, the participants additionally filled out an n-back strategy questionnaire that allowed us to examine the effect of reporting format (open-ended question vs. list-based query). The questionnaire provided several different strategies (e.g., rehearsal, updating, spatialization, guessing) together with descriptions. The participants were to assign a primary strategy, and optional secondary and tertiary strategies according to which strategy or strategies they had used in the last (15th) n-back block (see Appendix A). For the sake of comparison between rater categorizations (see below) and participant selections, the Grouping and Updating strategies in the strategy questionnaire were combined into a Grouping/Updating category, and the Semantic, Imagery, Spatialization, and Other strategy were combined into an Other category.

2.1.5. Rating participants' strategy descriptions

Independently of each other, two of the authors of this article classified each strategy report into one of seven different types. A strategy was defined liberally as the slightest hint of using some strategy, but a reiteration of the task instructions or a response that was unrelated to strategy use (e.g., only commenting on task difficulty) was not enough. See Table 2 for the strategy types together with response examples. If a participant reported multiple strategies for a single block, the most advanced strategy was used (see the strategies in Table 2 in ascending order, from less to more advanced; for empirical evidence supporting that this ranking is coupled with progressively higher performance on n-

Table 1
Descriptive information of the sample in Experiment 1.

Age	$M = 33.0$ ($SD = 8.4$), Range 18–50
Gender	Female 71.5%, Male 27.7%, Other 0.8%
Education	Lower secondary 3.8%
	Higher secondary 20.8%
	Basic vocational 10.8%
	Vocational university 10.8%
	Bachelor's degree 36.2%
	Master's degree 13.8%
	Doctoral degree 3.8%
MAI: Knowledge of cognition	$M = 3.71$, $SD = 0.58$, Range = 1.88–5.00
MAI: Regulation of cognition	$M = 3.55$, $SD = 0.49$, Range = 1.13–5.20
N-back: Average level of n	$M = 2.37$, $SD = 0.82$, Range = 2.27–4.64
N-back: Level of n in last block	$M = 2.91$, $SD = 1.37$, Range = 1–7
N-back: Strategy detail total score	$M = 27.92$, $SD = 18.32$, Range = 0–60

Note. $N = 130$. MAI = Metacognitive Awareness Inventory.

Table 2
Descriptions of the two strategy-related ratings: Strategy types and the level of detail in the applied strategy.

Strategy types	General description	Examples
No strategy	Using no strategy, empty response (and no prior response), irrelevant response, reiterating task instructions.	"No", "I had no strategy"
Guessing	Guessing, pressing response keys at random.	"I just randomly pressed one of the buttons when a number appeared", "I'm just guessing"
Familiarity	Relying on recognition memory, not actively trying to remember.	"I used my instinct", "Pressed according to how I felt I had seen"
Other	Very diverse category that contained strategies that did not fit into the other types, or descriptions that the raters could not categorize.	E.g., "I tried", "Relied on memory", "Ignore first digit and focus on the second", "I looked for a pattern"
Rehearsal	Repeating or rehearsing the items.	"Repeated the numbers in my head"
Grouping/ Updating	Grouping/chunking the items, or updating the items one by one or in chunks.	"I tried to repeat the last two numbers in my head", "Tried remembering 4 numbers in a row and then remember the next 4"
Grouping & comparison	Group/chunk and then compare the successive chunks to each other.	"I memorized a block of 3, then I examined the next three to see if they matched"
The level of detail in the strategy description		
0	No response, repeating task instructions, irrelevant	Empty (and no prior response), "No", "Difficult"
1	Very vague	"Yup", "I did my best"
2	Vague	"I used my memory"
3	General strategy	"I tried to say each number aloud"
4	General strategy including at least two details	"I memorized a block of 3, then I examined the next three to see if they matched"

[1] The pre-registration form did not include the last two exclusion criteria (previous experience with n-back, external help on n-back), but we felt confident that including them as exclusion criteria would improve data quality.

[2] Lures were defined as n+1 or n-1, i.e., adjacent to a possible match. In each block, two lures were n+1 and two were n-1. However, in the 2-back condition, due to the programming related to the automated block generator, the lures could simultaneously be n+1 and n-1 (e.g., 2-7-2-2). The 1-back lures only included n+1 lures as n-1 lures are not possible at this level of n. The target items could also simultaneously be lures (n+1 or n-1) if the level of n was long enough. E.g., in the 5-back block 5-5-1-7-5-9-5, the last digit 5 matches the item presented 5 items ago, but it also matches the item presented 6 items back (n+1 lure).

back tasks, see Laine et al., 2018; Forsberg et al., 2020). If a participant had reported a strategy on a previous block, but not responded to the question on the next one, the non-responded block was scored according to the most recent report. Besides strategy type, the raters also scored each report regarding the level of detail (0–4) of the written strategy report. A zero was marked if no response was given (with no prior responses that would have scored higher), if the response clearly stated that no strategy was used, if the response was a reiteration of the task instructions, or if the response did not concern the use of a strategy. A score of one entailed a very vague response that nevertheless gave some indication that the participant had tried to implement a strategy, while a somewhat less vague strategy-related response earned two points. A score of three required a description of a general strategy. For a score of four, the response had to convey a general strategy and at least two details of that strategy (see Table 2 for examples).

For the classification of strategy reports into the strategy types, the unweighted kappa ranged between 0.46 and 0.53 with an average of 0.50, which we deemed problematically low. However, upon closer inspection, we noted a systematic coding error for one of the raters. When this error was rectified (together with some haphazard additional errors that were observed during recoding), the kappa ranged between 0.65 and 0.78 with an average of 0.69, which we considered acceptable. For the scoring of the level of detail in the strategy reports, the weighted kappa ranged between 0.64 and 0.72 with an average of 0.69, which we again considered acceptable. The raters proceeded to perform consensus decisions for diverging strategy classifications and level of detail scores.

2.2. Results

2.2.1. Progress on the n-back task

The average n-back level per block is presented in Panel A of Fig. 1. The initial dip in performance is most probably explained by the fact that no practice trials were administered, and therefore many dropped down to the 1-back level after the initial 2-back block. Thereafter, a steady improvement is discernible, but there is considerable individual variation, as evidenced by the large standard deviations.

2.2.2. Emergence and stability of strategy use

According to the raters' classifications, approximately half of the participants (51.5%) reported using some kind of strategy already in the very first n-back block (see Panel B in Fig. 1). This number neared 100 (74.6%) towards the end of the task, but the steepest increase took place between the first and second blocks (see Panels B and C in Fig. 1). Out of the 130 participants, 26 (20%) were categorized as never using a strategy during the whole task. On average, the participants made 1.6 (SD = 1.9) strategy changes during the 15 n-back blocks (note that switching from and to no strategy was counted as a strategy change). About 40% (n = 52) did not change their reported strategy during the task (note that the 26 no-strategy users are included here). The rest of the participants changed their strategy once (n = 26), twice (n = 18), thrice (n = 12), or more than three times (n = 22). As depicted in Panel C of Fig. 1, most of the strategy changes took place during the first blocks, after which the change rates showed only a small decrease towards the end of the task. The level of detail in the written strategy reports ranged from 0 to 60, with an average of 27.9 (SD = 18.3). The strategy level of detail scores per block are depicted in Panel D of Fig. 1.

2.2.3. Strategy type and n-back performance

Four separate between-groups one-way ANOVAs (and two Welch's ANOVAs, see Table 3) were performed to test whether strategy type was associated with n-back performance. The strategy used in the last block served as the grouping factor, and the two separate dependent variables were the average n-back level over the 15 blocks and the level of n in the last block. Moreover, we ran separate analyses for the strategy type defined by ratings of the open-ended questions and the list-based query. For the analyses involving the list-based queries, the sample was slightly smaller (n = 106), as only those participants were included who had selected a single primary strategy. The last task block was chosen as the grouping factor as it was less ambiguous than, for example, the most commonly used strategy type; it was also the case that the list-based query was administered only at the end of the last block. All four analyses were statistically significant, indicating that strategy type was associated with objective n-back performance (see Table 3). Some strategy types were employed by only a limited number of participants but in all analyses, participants reporting Grouping and comparison and Grouping/updating exhibited highest mean levels of n-back performance, and the apparent superiority of these strategy types were also, to some degree, supported by post hoc pairwise comparisons (see Table 3).

2.2.4. Predictors of n-back performance

In a hierarchical multiple regression analysis model, we tested whether n-back performance was predicted by certain background factors (age, education, and metacognitive awareness) and strategy use as measured by the total level of detail in the open-ended strategy reports (see Table 1 for descriptives). The first step of this hierarchical multiple regression analysis model, including age, education, and the two metacognitive awareness measures as predictors, was non-significant. However, adding the total level of detail in the open-ended strategy reports at the second step yielded a statistically significant model, with the strategy level of detail variable explaining an additional 25.7% of the variance in n-back performance (Table 4). The bivariate correlations between all the variables are reported in Appendix B.

2.2.5. Metacognitive awareness and strategy development

The association between metacognitive awareness (MAI Knowledge of cognition and MAI Regulation of cognition) and strategy use in the n-back task was assessed in two ways. First, in two separate ANOVAs where strategy type was the grouping factor and MAI scores were the dependent variables (separate analyses were performed for each MAI variable), we observed no difference in MAI scores depending on what strategy the participants used in the third[3] n-back block (Knowledge of Cognition, $F(5, 124) = 1.62$, $p = .161$, $\eta^2 = 0.061$; Regulation of Cognition, $F(5, 124) = 1.70$, $p = .140$, $\eta^2 = 0.064$). Second, in a hierarchical multiple regression analysis controlling for age and education, the two MAI variables were not associated with the total strategy detail score in the n-back task, $\Delta F(2, 125) = 0.34$, $p = .72$, $\Delta R^2 = 0.005$. Hence, we found no associations between metacognitive awareness (as measured by the MAI) and strategy use in the n-back task (as measured by strategy type and level of detail in the open-ended strategy reports).

2.2.6. Comparison of the two strategy categorization methods (open-ended vs. list-based query) in the last task block

After completing the last n-back block, the participants responded not only to the open-ended strategy question, but also to the list-based strategy query. Comparison of the two query methods for the last task block revealed considerable discrepancy between the raters' classifications of the participants' open-ended strategy reports and the participants' list-based selection of their primary strategy (see Table 5). However, we also counted the number of participants whose strategy from the open-ended response in the 15th block matched any of the strategies they had indicated in the list-based query (irrespective of whether they marked it as a primary, secondary, or tertiary strategy). In this comparison, the congruency rate was 52.3%, and it also indicated that the discrepancy largely originated from those participants who were categorized as using an Other or No strategy on the basis of their open-ended response (see Table 6). At the same time, one should

[3] We chose not to use the first block as we speculated that at that point, many participants focused on understanding the task and remembering the response keys, as they received no training beforehand. We also ran the same analyses for the very last block with the same non-significant results.

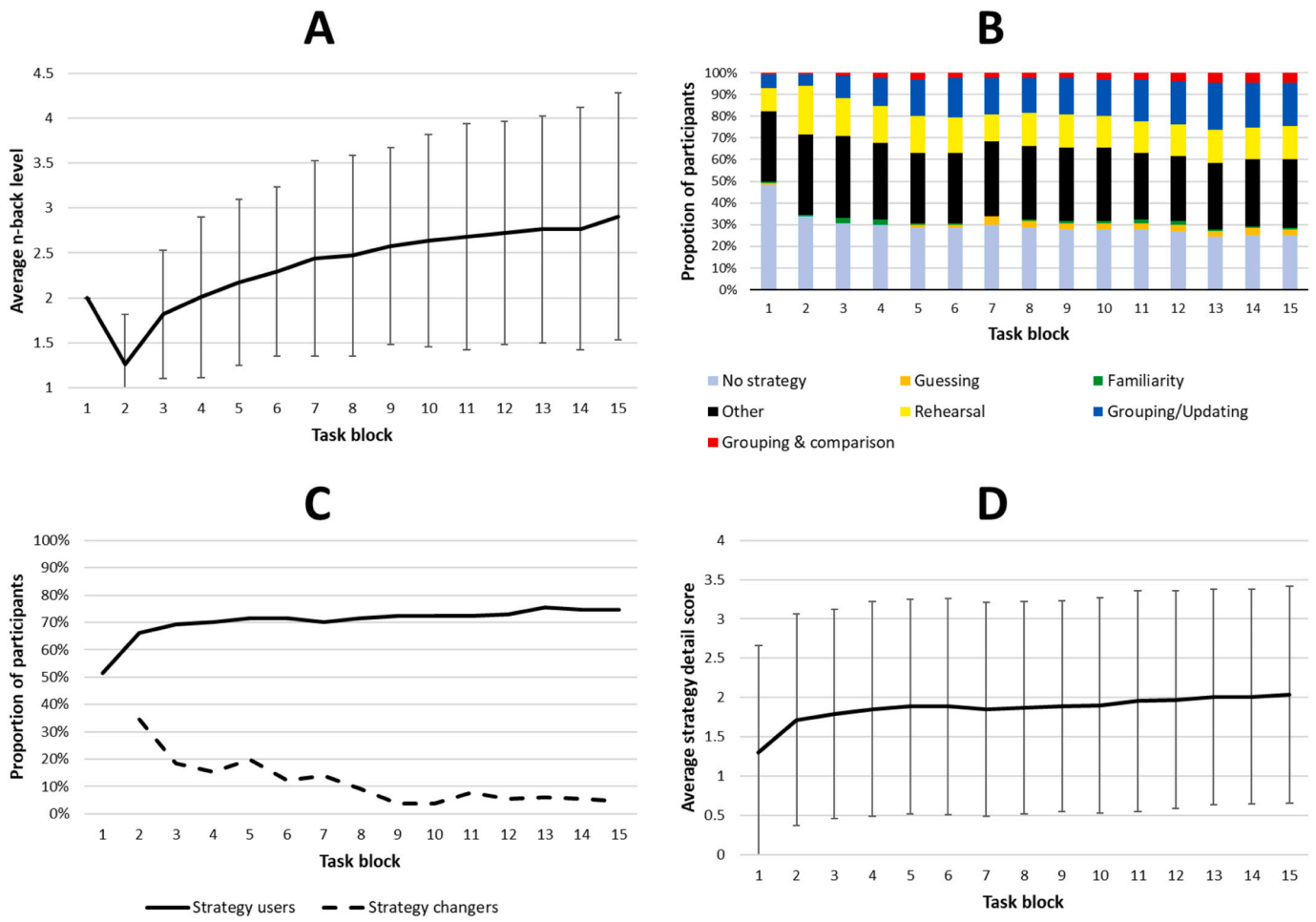


Fig. 1. Results of Experiment 1. $N = 130$ in all panels. *Panel A* - average n-back level in each task block. Error bars represent ± 1 *SD*. *Panel B* – strategy types employed per task block according to participants' open-ended strategy reports (see Table 3 for exact counts). The online version of this article contains bar colors. *Panel C* – strategy use and strategy change during the n-back task. The solid line represents the proportion of participants per task block who reported a strategy in their open-ended responses. The dashed line represents the proportion of participants who changed their strategy compared to the preceding task block. *Panel D* - average level of strategy detail per task block. Error bars represent ± 1 *SD*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

emphasize that another strategy category (Grouping/updating) showed quite high convergence. Importantly, strategy use was related to task success irrespective of the strategy query method (see Table 3).

2.3. Discussion

In our first experiment, we sought to elucidate how early participants exhibit strategy use in an n-back task and to what extent they stick to the strategy they select. We also examined whether strategy use (type of self-generated strategy and level of detail in its description) was associated with objective task performance.

As expected, the participants became better on the n-back task, on average, while showing large inter-individual variation (see panel A in Fig. 1). More importantly, our results indicate that over half of the participants reported strategy use already for the very first block of the n-back task. As most of the reported strategy changes also took place in the first task blocks, it seems that a substantial number of participants (but not all) adopted a strategy quite quickly and then stuck with it. These results fit well with the cognitive skill learning perspective that postulates fast strategy generation and implementation when faced with a novel task (Chein & Schneider, 2012; Taatgen, 2013).

Our results indicate that the type of strategy used was significantly associated with the level of performance on the n-back task. Looking at the mean values, the participants who reported using a grouping, updating, or grouping and comparison strategy performed best. Furthermore, the level of detail in the open-ended strategy reports explained an additional 25.7% of the variance in n-back performance after controlling for age, education, and metacognitive awareness. These results are in line with previous findings (Fellman et al., 2020; Forsberg et al., 2020; Laine et al., 2018) in showing that strategy use, coded from open-ended strategy reports, is significantly associated with performance on the n-back task. Thus, the present study extends the existing literature by showing that this association is evident already at the very first testing session.

With regard to remaining potential predictors of n-back performance, we did not find significant associations between age or education and n-back performance. Neither was metacognitive awareness, as assessed with the two MAI variables (Harrison & Vallin, 2017; Schraw & Dennison, 1994), significantly associated with n-back performance or strategy use (strategy type or level of detail in strategy reports). The lack of associations between the MAI and n-back may reflect the considerable differences between these measures. The items in the MAI seem more applicable to complex behaviors in everyday learning situations geared towards attaining non-immediate goals that require, for example, dividing an end-goal into sub-goals, organizing information, or gaining understanding (e.g., working as a student in a classroom). In contrast, the n-back is a very specific task with quite limited room for creativity, planning, or organizing.

Table 3
One-way ANOVAs and descriptives on the relationship between strategy type and n-back performance.

	n	*M*	*SD*	Min	Max
Strategy type in the last block based on open-ended question: Average n-back level across 15 blocks $F(5, 123) = 11.21$, $p < .001$, $\eta^2 = .313$[a]					
No	33	1.88	0.48	1.13	2.93
Guessing	3	2.29	0.20	2.07	2.47
Familiarity	1	2.53			
Other	41	2.28	0.85	1.20	5.20
Rehearsal	20	2.37	0.33	1.53	3.07
Grouping / Updating	26	2.75	0.87	1.27	5.07
Grouping & comparison	6	3.94	0.57	3.13	4.67
Strategy type in the last block based on open-ended question: Level of n in the last block $F(5, 123) = 9.67$, $p < .001$, $\eta^2 = .282$[b]					
No	33	2.09	0.98	1	5
Guessing	3	3.00	1.00	2	4
Familiarity	1	3.00			
Other	41	2.63	1.41	1	7
Rehearsal	20	3.15	0.75	2	5
Grouping / Updating	26	3.69	1.41	1	7
Grouping & comparison	6	5.00	0.63	4	6
Strategy type in the last block based on list-based query: Average n-back level across 15 blocks $F(5, 100) = 2.39$, $p = .043$, $\eta^2 = .107$[c]					
No	2	2.20	0.19	2.07	2.33
Guessing	0				
Familiarity	16	2.16	0.47	1.53	3.13
Other	7	2.49	1.27	1.53	5.20
Rehearsal	38	2.26	0.79	1.20	5.07
Grouping / Updating	31	2.73	0.76	1.80	4.67
Grouping & comparison	12	2.88	0.86	1.33	4.40
Strategy type in the last block based on list-based query: Level of n in the last block $F(5, 100) = 3.33$, $p = .008$, $\eta^2 = .143$[d]					
No	2	3.50	0.71	3	4
Guessing	0				
Familiarity	16	2.69	1.20	1	5
Other	7	3.00	1.53	1	6
Rehearsal	38	2.55	1.20	1	6
Grouping / Updating	31	3.61	1.52	1	7
Grouping & comparison	12	3.83	1.19	2	5

Note. $N = 130$ for the analyses involving the rater-based strategy categorizations. $n = 106$ for the analyses involving the strategy types from the list-based query.

[a] The single participant using the Familiarity strategy was removed ($n = 129$). As the assumption of homoscedasticity was violated, a Welch's ANOVA was also performed. This analysis was also significant, $F(18.06) = 14.40$, $p < .001$, indicating that there were some significant differences in n-back performance between the strategy types. Post hoc comparisons using Games-Howell test indicated that the Grouping and comparison strategy was significantly better than all other strategy types (p-values $< .05$); and that the Grouping/Updating and Rehearsal strategies were significantly better than No strategy ($p < .05$).

[b] The single participant using the Familiarity strategy was removed ($n = 129$). As the assumption of homoscedasticity was violated, a Welch's ANOVA was also performed. This analysis was also significant, $F(16.18) = 16.83$, $p < .001$, indicating that there were some significant differences in n-back performance between the strategy types. Post hoc comparisons using Games-Howell test indicated that the Grouping and comparison strategy was significantly better than all (p-values $< .05$) but the Guessing strategy ($p = .221$); Grouping/Updating was significantly better than the Other and No strategy types (p-values $< .05$); and Rehearsal was significantly better than No strategy ($p = .001$).

[c] Post hoc comparisons using Tukey HSD test indicated no statistical pairwise differences.

[d] Post hoc comparisons using Tukey HSD test indicated that the Grouping and comparison and Grouping/Updating strategy categories were significantly better than the Rehearsal strategy (p = .047 and .016 respectively). No other pairwise comparisons were statistically significant.

Experiment 1 had some limitations that are important to point out. One fifth of the participants never implemented a strategy according to the raters' categorizations of the open-ended responses. However, according to the list-based query responses, only ca 2% of the participants reported using no strategy in the final task block. Although this particular discrepancy largely reflects non-responding on the open-ended queries, it also applies to some extent to the other types of strategies (Table 5). Despite this discrepancy, the statistical analyses showed that strategy type was significantly associated with n-back performance irrespective of query method (open-ended or list-based).[4] However, both raters experienced it challenging to categorize and score several of the open-ended strategy reports, which might have introduced a potential source of subjective bias that could have affected the results. Additional issues related to potential rater bias are that the raters were aware of the aims of this study, which responses stemmed from the same participant, and the order of the responses; and that some open-ended responses explicitly mentioned the level of n, which might have inadvertently affected the raters. A less ambiguous alternative to rating open-ended strategy reports would be to use the list-based query format to follow up strategy use in a WM task (see e.g. Dunlosky & Kane, 2007; Wu et al., 2008). However, this method runs the risk of modulating subsequent strategy use, as it provides the participant with information on the major strategy choices. To examine whether this is the case, we ran a second n-back experiment where we randomized participants into an open-ended strategy report group and a list-based query group.

Table 4
Hierarchical regression analysis for variables predicting average n-back level.

Predictor		N-back: Average level of n			
		ΔF	ΔR^2	β	*B*
Step 1		1.53	.047		
	Age			−.009	−.001
	Education			.199*	.102
	MAI: Knowledge of cognition			.038	.054
	MAI: Regulation of cognition			.023	.038
Step 2		45.64***	.257		
	Age			.052	.005
	Education			.115	.059
	MAI: Knowledge of cognition			.082	.115
	MAI: Regulation of cognition			−.003	−.006
	Strategy detail sum score			.517***	.023

Note. $N = 130$. MAI = Metacognitive Awareness Inventory.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

3. Experiment 2

The second experiment was an adaptation of Experiment 1, including a pretest questionnaire, an adaptive n-back task, and a posttest questionnaire. One main difference was the inclusion of two groups. One group gave open-ended strategy reports after each n-back block (identical to Experiment 1), while the other group responded to a list-based strategy query after each block. This allowed us to achieve our two main aims: replicating the findings from Experiment 1 and examining whether administration of list-based strategy queries after each block affects participants' self-reported use of strategies and/or their progress in the n-back task. Another major change was replacing the MAI with the Memory Aids Questionnaire (MAQ; Chouliara & Lincoln, 2015) to see if another potentially significant predictor, self-reported use of internal and/or external memory aids in everyday life, is associated with strategy use and/or performance on the n-back task. Especially more frequent employment of internal memory aids in everyday life might reflect higher proneness to apply strategies when faced with a novel memory task. Furthermore, as it is not fully clear which aspects of cognition the open-ended strategy reports reflect, we added a simple picture-

[4] However, if the most advanced list-based strategy (akin to how the open-ended responses were coded) was used as the grouping factor, i.e., irrespective of whether it had been reported as a primary, secondary, or tertiary strategy, the effects were no longer significant in the list-based comparison, $p = .056$ and 0.059. Concern has, however, been raised regarding the validity and reliability of the reported "lesser" strategies (Morrison et al., 2016).

Table 5
Cross-tabulation of the categorizations of open-ended strategy reports (horizontal) vs. list-based query responses (vertical) on the 15th n-back block. The query responses reflect the selected primary strategy.

	None	Guessing	Familiarity	Other	Rehearsal	Gr/Up	Gr & comp	Rater total
None	1	0	10	1	9	1	3	25
Guessing	1	0	1	0	1	0	0	3
Familiarity	0	0	0	0	1	0	0	1
Other	0	0	1	2	15	9	2	29
Rehearsal	0	0	2	4	7	6	0	19
Gr/Up	0	0	2	0	5	13	3	23
Gr & comp	0	0	0	0	0	2	4	6
Participant total	2	0	16	7	38	31	12	

Note. Gr = Grouping, Up = Updating, comp = comparison. Only participants who had selected a single primary strategy in the strategy questionnaire (excluding "Did not understand", see Appendix A) were included ($n = 106$). For the sake of comparison, the following strategies in the list-based query have been combined: Grouping, Updating = Grouping/Updating; Semantic, Imagery, Spatialization, Other strategy = Other.

Table 6
Number of participants whose strategy categorization from the open-ended response in the 15th block matched any of the strategies they had indicated in the list-based query (irrespective of whether they marked it as primary, secondary, or tertiary).

	No	Guessing	Familiarity	Other	Rehearsal	Gr/Up	Gr & comp
				Experiment 1			
No-match	19	2	1	30	5	3	2
Match	14	1	0	11	15	23	4
				Experiment 2			
No-match	15	2	0	15	3	2	0
Match	9	1	4	2	8	10	3

Note. No = No strategy, Gr/Up = Grouping/Updating, Gr & comp = Grouping and comparison.

description task in order to explore whether a more general cognitive-linguistic feature, verbal productivity, is associated with the strategy reports and n-back performance. As noted above, it seems feasible to assume that the employment of a strategy, a consciously chosen verbalizable method used to perform a task, is strongly guided by the language system. Finally, we probed the development of task routine through perceived response key mastery and tested if it is associated with n-back performance. For this purpose, we added a question after each n-back block. Detailed descriptions of these changes and some additional minor changes are described in the Methods. The preregistration of this experiment can be found at http://aspredicted.org/blind.php?x=gx5g9t.

3.1. Methods

3.1.1. Ethics statement

The experiment was approved by the joint ethics committee of the Departments of Psychology and Logopedics at Åbo Akademi University. We obtained informed consent from all participants, participation was anonymous, and we informed participants of their right to withdraw from the experiment at any time.

3.1.2. Procedure

The basic procedure in Experiment 2 was identical to that of Experiment 1: the online data collection encompassed a pretest questionnaire, an adaptive n-back task, and a posttest questionnaire. The pretest questionnaire consisted of the same background questionnaire as in Experiment 1, with the addition of a simple picture description task. In the picture description task, the participants described a weather-photograph with as much detail as possible. The picture depicted a cloudy sky that was pierced by some bright yellow sunlight, and in the very bottom of the picture, top branches of trees were seen. We used the number of generated words (irrespective of content) as the dependent variable for this task. Compared to Experiment 1, the main difference in Experiment 2 was the inclusion of two separate groups that reported their use of strategies in the n-back task differently. One group gave list-based strategy reports while the other gave open-ended strategy reports throughout the task blocks (see below for further details). Participants were randomly allocated to the two groups. The post-task questionnaire consisted of the MAQ and the same posttest questionnaire as in Experiment 1. The only exception was that the list-based strategy group received an additional question related to whether the list of strategy alternatives had affected their use of strategies in the n-back task (Yes/No response).

3.1.3. Participants

We recruited participants in three batches using the Prolific crowdworking site. For the first two batches (10 + 130 participants), we invited a sample of participants that we had previously pre-screened for another study and who fulfilled most of our inclusion criteria. As this did not provide our minimum requirement of 60 participants per group, we collected data from a third batch (50 participants) that had not been previously pre-screened. For the third batch, we used Prolific's built-in prescreening tool to target participants who were 18–51 years old, as this matched the age range for batches 1–2. Altogether 187 participants completed Experiment 2: 89 in the list-based strategy reports group and 98 in the open-ended strategy reports group. Ultimately, 63 participants in the list-based strategy reports group and 75 participants in the open-ended strategy reports group fulfilled our pre-registered inclusion criteria[5] (see Table 7 for descriptives). However, we additionally excluded one participant in the open-ended strategy reports group who was an extreme outlier by having attained the 13-back level (the next closest participant having reached 7-back), one participant in the list-

[5] The inclusion criteria based on self-report were as follows: no neurological or psychiatric illness that affects the life of the participant, no specific learning disabilities (e.g., language disorders, attention disorder), no use of medication or drugs affecting the CNS (except tobacco, alcohol, and marijuana), not being intoxicated at the time of testing, consuming less than 10 units of alcohol on the day before testing, reaching at least the 2-back level during the 15 n-back blocks, no previous n-back experience, no use of external help (e.g., note-taking) while completing the n-back task.

Table 7
Descriptive information of the two groups in Experiment 2.

		List-based strategy reports group ($n = 61$)	Open-ended strategy reports group ($n = 74$)
Age		$M = 31.3$, $SD = 8.4$	$M = 35.1$, $SD = 9.0$
Gender		50.8% female, 49.2% male	63.5% female, 36.5% male
Education	Primary	0%	1.4%
	Lower secondary	0%	0%
	Higher secondary	29.5%	24.3%
	Basic vocational	8.2%	10.8%
	Vocational university	6.6%	9.5%
	Bachelor's degree	39.3%	41.9%
	Master's degree	16.4%	12.2%
	Doctoral degree	0%	0%
MAQ: Internal		$M = 12.00$, $SD = 2.71$, Range = 5–16[a]	$M = 10.77$, $SD = 3.76$, Range = 0–16
MAQ: External		$M = 12.88$, $SD = 2.66$, Range = 6–16[a]	$M = 12.41$, $SD = 2.99$, Range = 0–16
Picture description task: Word count		$M = 36.20$, $SD = 28.23$, Range = 4–193	$M = 31.07$, $SD = 19.16$, Range = 4–91
N-back: Average level of n		$M = 2.85$, $SD = 0.94$, Range = 1.13–5.40	$M = 2.43$, $SD = 0.83$, Range = 1.13–4.60
N-back: Level of n in last block		$M = 3.54$, $SD = 1.59$, Range = 1–8	$M = 2.93$, $SD = 1.56$, Range = 1–7
N-back: Strategy detail total score		N/A	$M = 25.70$, $SD = 19.62$, Range = 0–56
N-back: Task routinization score		$M = 87.38$, $SD = 40.50$, Range = 16–147	$M = 75.07$, $SD = 38.12$, Range = 19–142

Note. MAQ = Memory Aids Questionnaire.
[a] One participant reported not using internal or external memory aids, hence $n = 60$.

based strategy reports group whose user account had been banned by Prolific for multiple account ownership, and one participant in the list-based group who had completed our study twice (i.e., the data from the second participation was removed), even though we had not listed these exclusion criteria in our pre-registration. Hence, $n = 61$ and 74 for the list-based strategy report group and the open-ended strategy reports group, respectively.

According to chi-square tests, the groups did not differ significantly on gender distribution, $\chi(1) = 2.21$, $p = .14$, or education level, $\chi(5) = 2.22$, $p = .82$. However, an independent samples *t*-test indicated that the list-based strategy reports group was significantly younger than the open-ended strategy reports group, $t(133) = 2.51$, $p = .01$.

3.1.4. The adaptive n-back task

The adaptive n-back task in Experiment 2 was the same as in Experiment 1, except for the following differences:

(1) We used two task variants, and each group completed only one of them. In one variant, the participants gave list-based strategy reports after each of the 15 n-back blocks. The list used the same strategies and descriptions as in the posttest questionnaire in Experiment 1 (Appendix A), but instead of simultaneously defining a primary, secondary, and tertiary strategy (as in Experiment 1), the participants first selected a single primary strategy and then, on a separate page, any secondary strategies. This change was made because several participants had selected multiple primary strategies in Experiment 1. The second task variant where the participants gave open-ended strategy reports was identical to that employed in Experiment 1.

We used two task variants, and each group completed only one of them. In one variant, the participants gave list-based strategy reports after each of the 15 n-back blocks. The list used the same strategies and descriptions as in the posttest questionnaire in Experiment 1 (Appendix A), but instead of simultaneously defining a primary, secondary, and tertiary strategy (as in Experiment 1), the participants first selected a single primary strategy and then, on a separate page, any secondary strategies. This change was made because several participants had selected multiple primary strategies in Experiment 1. The second task variant where the participants gave open-ended strategy reports was identical to that employed in Experiment 1.

(2) After each n-back block, the participants rated on a scale from 1 (Very difficult) to 10 (very easy/automatic) how easy the responding felt ("In the previous n-back sequence, how easy was it to give a response without thinking about which is the 'Same' and which is the 'Not same' button?"). This rating aimed to address automatization of one aspect of task performance. In addition, after each n-back block, the participants rated on a scale from 1 (Not at all) to 10 (Extremely) how mentally demanding the block had felt.

(3) Due to participant feedback in Experiment 1, we added the completed block number to the results screen (e.g., 7/15), which enabled the participants to keep track of their progress.

3.1.5. Rating participants' strategy descriptions

Independently of each other, the same two raters as in Experiment 1 categorized each open-ended strategy report and scored them on their level of detail. For the classification of strategy reports into strategy types, the unweighted kappa ranged between 0.77 and 0.83 with an average of 0.80, which we deemed adequate. For the scoring of the level of detail in the strategy reports, the weighted kappa ranged between 0.82 and 0.86 with an average of 0.84, which again was considered adequate. Therefore, the raters proceeded to perform consensus decisions for diverging ratings.

3.2. Results

3.2.1. Replication of experiment 1

For the following analyses, we have only included the open-ended strategy reports group, as it corresponds to Experiment 1 regarding strategy reporting. A second reason for excluding the list-based strategy group from these analyses was our finding that the list-based queries facilitated strategy use and n-back performance (see Section 3.2.3 below).

3.2.1.1. Progress on the n-back task. The results on task progress are in line with those obtained in Experiment 1 (compare panels A in Figs. 1 & 2).

3.2.1.2. Emergence and stability of strategy use. The strategy-related results generally replicated those of Experiment 1 (Fig. 2). According to the raters' categorizations, approximately half of the participants (52.7%) reported using some kind of strategy already in the very first n-back block (see Panel B in Fig. 2). In the last block, 67.6% were categorized as implementing some kind of strategy. As in Experiment 1, the steepest increase in strategy use took place between the first and second blocks (see Panels B and C in Fig. 2). Out of the 74 participants, 19 (25.7%) were categorized as never using a strategy. On average, the participants made 1.4 ($SD = 1.6$) strategy changes during the task. The majority of participants never changed their strategy ($n = 30$; note that the 19 no-strategy users are included here). Fewer participants changed their strategy once ($n = 15$), twice ($n = 13$), thrice ($n = 6$), or more than three times ($n = 10$). As depicted in Panel C of Fig. 2, most of the strategy changes happened during the first blocks, after which the rates of strategy changes dropped and remained at more or less the same levels up to the end of the task. The sum score of the level of detail in the

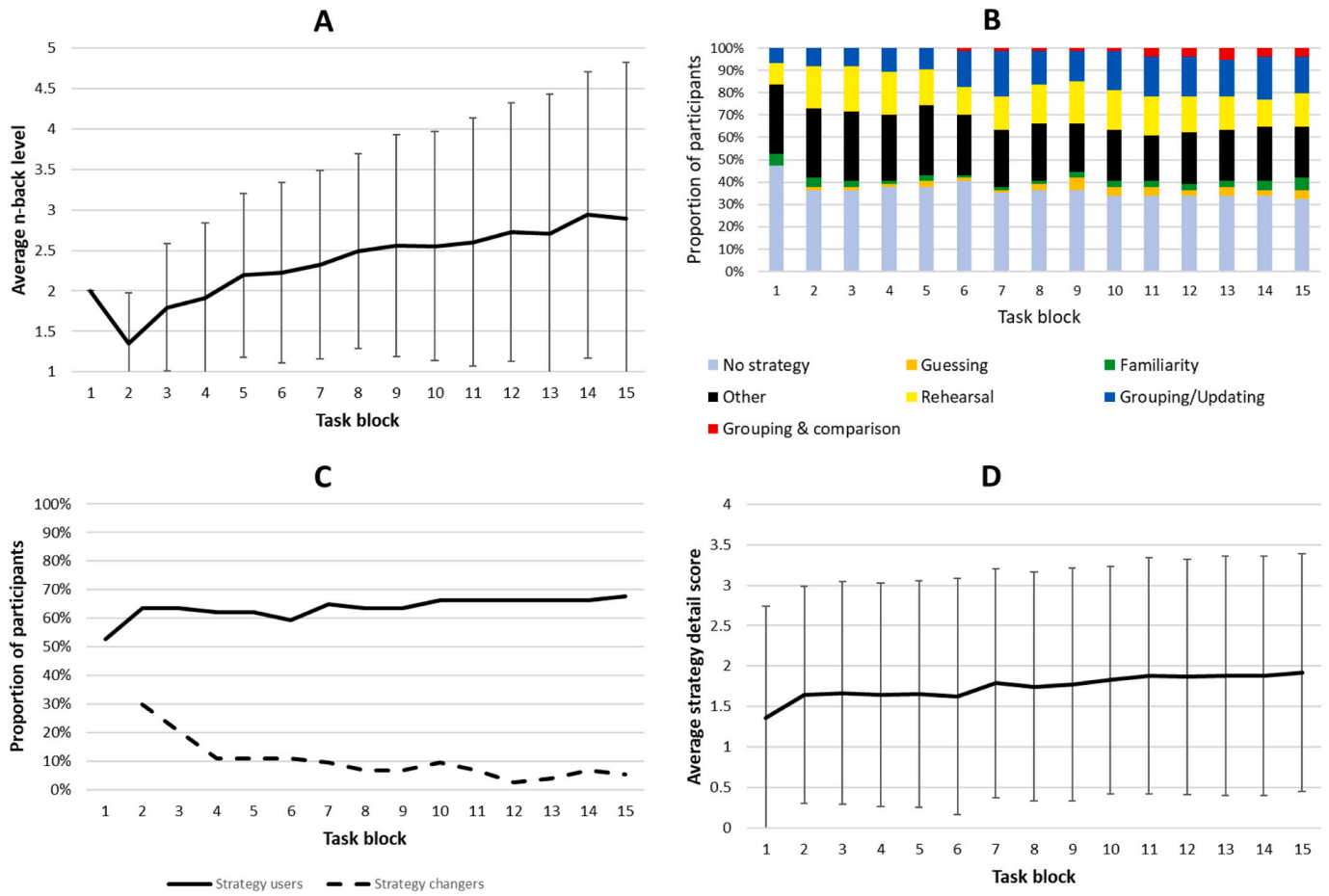


Fig. 2. Results of Experiment 2. $N = 74$ in all panels. *Panel A* - average n-back level in each block of the task, error bars represent ± 1 *SD*. *Panel B* - the raters' categorizations of the participants' open-ended strategy reports. The online version of this article contains bar colors *Panel C* - the solid line represents, per block, the proportion of participants who, according to the raters' categorizations of the participants' open-ended responses, used a strategy; the dashed line represents the proportion of participants who changed their strategy from the first block to the second, from second to third etc. *Panel D* - average level of strategy detail score in each block, error bars represent ± 1 *SD*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

written strategy reports ranged from 0 to 56, with an average of 25.7 ($SD = 19.6$). The strategy level of detail scores per block are depicted in Panel D of Fig. 2.

3.2.1.3. Strategy type and n-back performance. We performed identical ANOVAs as in Experiment 1 and again obtained significant results in all but one of the Welch's ANOVAs (Table 8). Thus, these results generally replicated Experiment 1 in showing that strategy type was significantly related to objective n-back performance. Again, Grouping & comparison and Grouping/updating were associated with the highest average n-back performance levels, but post hoc comparisons only indicated a significant advantage for the Grouping & comparison strategy type.

3.2.2. Predictors of n-back performance

We tested with two hierarchical multiple regression analysis models whether n-back performance level was predicted by certain factors (see Table 7 for descriptives). In the first model that was akin to the one in Experiment 1, the first step with age and education as predictors was non-significant. However, adding the average level of detail in the open-ended strategy reports at the second step yielded a statistically significant model, with the strategy level of detail variable explaining an additional 9.9% of the variance in n-back performance (Table 9). The bivariate correlations between all the variables are reported in Appendix B. In the second regression analysis model, the first step, including the two MAQ composite scores (internal & external memory aids, 4 items each), the n-back task routinization sum score, and the verbal productivity score (picture description word count), yielded a statistically significant model that accounted for 12.7% of the variance in n-back performance. Of the four predictors, verbal productivity was the only statistically significant predictor (Table 10). Adding the average level of detail in the open-ended strategy reports at the second step yielded a statistically significant model that explained an additional 6.2% of the variance in n-back performance. In this model, the level of strategy detail was the only significant predictor. This would suggest that the verbal productivity and strategy description scores show some overlap regarding the variance they explain in n-back performance (their bivariate correlation was $r = 0.32$, $p = .005$).

3.2.3. The list-based strategy reports group: n-back performance and strategy use

With the following analyses, we tested whether the list-based and open-ended strategy reports groups differed concerning n-back performance and strategy use, as it would indicate whether the repeated presentation of the strategy list had affected the participants in the list-based strategy reports group. Concerning the average n-back level across the 15 task blocks, an independent samples t-test indicated that the list-based strategy reports group ($M = 2.85$, $SD = 0.94$) performed significantly better than the open-ended strategy reports group ($M = 2.43$, $SD = 0.83$), $t(133) = 2.79$, $p = .006$, $d = 0.48$. Out of the 61 participants in the list-based strategy reports group, 32 (52.5%) reported that the

Table 8
One-way ANOVAs and descriptives on the relationship between strategy type and n-back performance.

	n	*M*	*SD*	Min	Max
Strategy type in the last block based on open-ended question: Average n-back level across 15 blocks $F(6, 67) = 2.51$, $p = .030$, $\eta^2 = .184$[a]					
No	24	2.12	0.72	1.13	3.73
Guessing	3	2.62	1.37	1.33	4.07
Familiarity	4	2.82	0.50	2.27	3.47
Other	17	2.16	0.48	1.33	3.07
Rehearsal	11	2.55	1.11	1.33	4.60
Grouping / Updating	12	2.94	0.83	1.53	3.93
Grouping & comparison	3	3.20	0.82	2.60	4.13
Strategy type in the last block based on open-ended question: Level of n in the last block $F(6, 67) = 4.31$, $p = .001$, $\eta^2 = .278$[b]					
No	24	2.21	1.10	1	5
Guessing	3	3.67	3.06	1	7
Familiarity	4	3.25	1.26	2	5
Other	17	2.47	0.94	1	4
Rehearsal	11	3.36	1.80	1	7
Grouping / Updating	12	3.67	1.61	1	7
Grouping & comparison	3	5.67	0.58	5	6
Strategy type in the last block based on list-based query: Average n-back level across 15 blocks $F(5, 68) = 5.49$, $p < .001$, $\eta^2 = .288$[c]					
No	5	1.95	0.57	1.33	2.60
Guessing	0				
Familiarity	15	2.24	0.68	1.33	3.67
Other	6	1.96	0.80	1.33	3.07
Rehearsal	20	2.12	0.72	1.33	3.93
Grouping / Updating	19	2.72	0.80	1.53	4.60
Grouping & comparison	9	3.38	0.66	2.27	4.13
Strategy type in the last block based on list-based query: Level of n in the last block $F(5, 68) = 9.10$, $p < .001$, $\eta^2 = .401$[d]					
No	5	1.80	0.84	1	3
Guessing	0				
Familiarity	15	2.40	1.12	1	5
Other	6	1.83	1.17	1	4
Rehearsal	20	2.55	1.10	1	5
Grouping / Updating	19	3.32	1.42	1	7
Grouping & comparison	9	5.22	1.56	2	7

Note. $N = 74$ in all analyses (i.e., the open-ended strategy reports group).

[a] As the assumption of homoscedasticity was violated, a Welch's ANOVA was also performed. This analysis was non-significant, $F(11.09) = 2.34$, $p = .105$, indicating that there were no significant differences in n-back performance between the strategy types.

[b] As the assumption of homoscedasticity was violated, a Welch's ANOVA was also performed. This analysis was significant, $F(11.66) = 10.93$, $p < .001$, indicating that there were some significant differences in n-back performance between the strategy types. Post hoc comparisons using Games-Howell test indicated that the Grouping and comparison strategy was significantly better than the Rehearsal, Other, and No strategy categories p-values $< .05$). No other pairwise comparisons were statistically significant.

[c] Post hoc comparisons using Tukey HSD test indicated that the Grouping and comparison strategy was significantly better than the Rehearsal, Other, Familiarity, and No strategy types (p-values $< .05$). No other pairwise comparisons were statistically significant.

[d] Post hoc comparisons using Tukey HSD test indicated that the Grouping and comparison strategy was significantly better than all other strategies (p-values $< .01$). No other pairwise comparisons were statistically significant.

Table 9
Hierarchical regression analysis for variables predicting average n-back level across 15 n-back blocks.

Predictor		N-back: Average level of n			
		ΔF	ΔR^2	β	*B*
Step 1		2.63	.069		
	Age			−.082	−.008
	Education			.246*	.137
Step 2		8.34**	.099		
	Age			−.016	−.001
	Education			.153	.085
	Strategy detail sum score			.336**	.014

Note. $N = 74$.
* $p < .05$.
** $p < .01$.

Table 10
Hierarchical regression analysis for variables predicting average n-back level across 15 n-back blocks.

Predictor		N-back: Average level of n			
		ΔF	ΔR^2	β	*B*
Step 1		2.51*	.127		
	MAQ: Internal			−.124	−.027
	MAQ: External			.016	.004
	N-back task routinization			.200	.004
	Verbal productivity			.255*	.011
Step 2		5.18*	.062		
	MAQ: Internal			−.110	−.024
	MAQ: External			.020	.006
	N-back task routinization			.096	.002
	Verbal productivity			.167	.007
	Strategy detail sum score			.285*	.012

Note. $N = 74$. MAQ = Memory Aids Questionnaire.
* $p < .05$.

presentation of the strategy list at the end of each n-back block had affected their strategy use. A follow-up ANOVA with a Tukey post hoc test indicated that the subgroup of participants who reported being affected by the list ($n = 32$, $M = 3.11$, $SD = 1.01$) performed significantly better than the open-ended strategy reports group ($p = .001$) and the subgroup of participants reporting not being affected by the list ($n = 29$, $M = 2.57$, $SD = 0.77$, $p = .04$), while there was no statistically significant difference between the subgroup of participants who reported not being affected by the list and the open-ended strategy reports group ($p = .739$, see Fig. 3). Furthermore, a chi-square test also indicated that the subgroup that reported that they had been influenced by the list reported using more advanced (see Table 2) primary strategies in any of the 15 blocks than the subgroup reporting not being affected by the list, $\chi^2(4, N = 61) = 11.984$, $p = .017$. However, a separate chi-square test on the last block indicated no significant differences in list-based strategy sophistication between the subgroups within the list-based strategy reports group and the open-ended strategy reports group, $\chi^2(12, N = 135) = 17.524$, $p = .131$.

All in all, these results indicate that exposure to the list of strategy descriptions affected n-back performance. Additional analyses suggested that this effect was driven by individuals who reported that they had been affected by the strategy list. Moreover, there was some evidence for a more common use of sophisticated strategies by this subgroup, which could be related to their better n-back performance. Because of this potential confound, further analysis of the list-based strategy reports group's strategy use is problematic, but we have nevertheless included more detailed descriptions of the two subgroups' strategy reports in Appendix C.

3.3. Discussion

The results of the open-ended strategy reports group in our second experiment generally replicated the main findings of Experiment 1: strategies were adopted early on, the highest rates of changes in strategy were observed in the first blocks, and strategy level of detail and strategy type were related to n-back performance level (with the grouping/updating, and Grouping and comparison strategies showing the highest mean values, although post hoc comparisons only indicated superiority of the Grouping and comparison strategy). These replications provide further support to the cognitive skill learning view (Chein & Schneider, 2012; Taatgen, 2013) on WM task performance. The variance in n-back

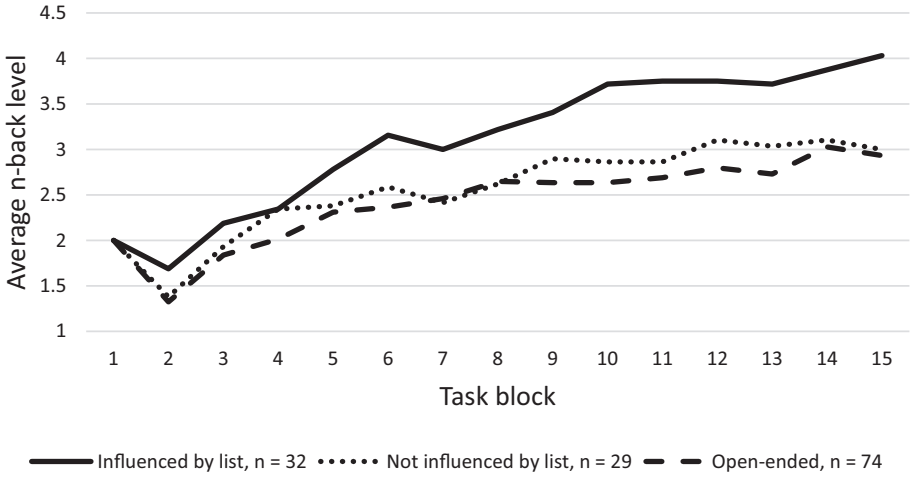


Fig. 3. Average n-back level in each block of the n-back task for the open-ended strategy reports group and the two subgroups identified in the list-based strategy reports group.

performance that was explained by the level of strategy detail was noticeably lower in Experiment 2 (9.9% vs. 26.7%), which could reflect random variance due to sample size. For instance, Schönbrodt and Perugini (2013) suggest that correlations stabilize when sample sizes approach 250, and thus there is likely more random variance in Experiment 2 ($n = 74$) than in Experiment 1 ($n = 131$).

Our second main aim in Experiment 2 was to test whether administering list-based strategy queries after each n-back block would affect participants' strategy adoption and/or objective performance. Our results showed that the list-based query format had in fact affected participants' behavior. Approximately half of the participants in the list-based strategy query group reported that they had picked up a strategy from the list, and these participants used more sophisticated strategies and performed significantly better on the n-back task than the open-ended strategy reports group. Thus, these results raise concern for using this kind of query format multiple times in a test battery or in a follow-up, as it could influence participants' task performance.

Concerning predictors of n-back performance, the MAQ internal and external memory aids variables (Chouliara & Lincoln, 2015) were unrelated to n-back performance in the current experiment. This could be because the MAQ was intended as an outcome questionnaire on the effectiveness of rehabilitation for neurological patients, and we assessed neurologically intact individuals. On the other hand, akin to our null findings regarding the MAI in Experiment 1, questionnaires related to real-world behaviors and activities may not be directly applicable for a rather abstract cognitive test like the n-back. Our simple task routinization measure (how automatic the key mapping and responding had felt) was not significantly associated with n-back performance. However, it is possible that several participants misinterpreted this question, as the within-individual ratings fluctuated quite a bit, which appears counterintuitive. Perhaps they interpreted the question so that it also probed how well they thought they had performed on a given task block. Finally, verbal productivity, here measured by a word count on a simple written picture description task, was significantly associated with n-back performance. Interestingly, verbal productivity and the level of detail in the strategy reports were significantly correlated ($r = 0.32$) and partly overlapped as predictors of n-back performance. This suggests that, to some degree, the open-ended strategy reports reflect verbal abilities. This was not unexpected given our assumption that the use of strategies that we have defined as conscious, verbalizable ways to handle a task would be strongly guided by the language system. It remains open to what extent also factors like general intelligence and motivation (the effort put on the tasks) underlie the results on these measures. In our current experiment, verbal productivity did not correlate with a measure of self-rated motivation ($r = -0.05$, see Appendix B), but one should note that the motivation measure was significantly negatively skewed.[6]

4. General discussion

Using the cognitive skill learning perspective (Chein & Schneider, 2012; Taatgen, 2013) as our general framework, we ran a dual experiment study to examine the hitherto unexplored block-by-block development of strategy use in a widely employed WM updating task (n-back) that was novel to the research participants. More specifically, we were interested in finding out how quickly participants develop strategies for the n-back task and whether they stick to a specific strategy, how effective strategy use is in terms of objective n-back performance, and whether selected predictors are related to strategy development and n-back performance. We assumed that strategy implementation could be quite fast in a task like the n-back that is rather straightforward, and that strategy use would stabilize when the task becomes more familiar. An additional methodological aim was to test whether within-task list-based strategy queries affect strategy use and task performance.

Concerning our first set of aims, both experiments indicated that more than half of our participants adopted strategies already during the 20 first items of the n-back task. Thus, this happened within the first 1–2 min into performing a novel WM task that the participants had never practiced or seen before. Based on the list-based strategy queries in Experiment 2, this percentage was even higher (81.7%), but we cannot exclude the possibility that their reports were affected by the exposure to the list. As regards changes in strategy, the highest rates of strategy changes were observed during the first task blocks. After that, the rates dropped to lower levels towards the end of the task (note, however, that this decline was not evident for one of the subgroups in the list-based strategy reports group, see Fig. C4 in Appendix C).

Overall, the present results show that the initial stages in performing a novel WM task are very dynamic in terms of strategy use, which fits well with the cognitive skill learning view that implies strategy generation in the initial stages of task learning (Chein & Schneider, 2012; Taatgen, 2013). We did not have specific hypotheses about the exact time-course of strategy generation, as the skill learning framework is general and has not been previously applied to the n-back task. Rather,

[6] Shapiro-Wilk test of normality, $W(74) = 0.82$, $p < .001$. Motivation $M = 8.6$, $SD = 1.6$, skewness $= -1.34$, kurtosis $= 1.99$ (approximately 81% had given a self-rating of 8 or higher on a scale from 1 to 10, where 10 indicated Very motivated).

the present results provide descriptive data about the time course of strategy generation in this type of task.

One should note that despite its generality, the skill learning framework as presented by Chein and Schneider (2012) may not be directly applicable to all task situations. First, their model implies not only that the metacognitive phase starts early, but also that it stops early as the processing shifts to the controlled execution and automation phases: "[…] humans can learn to perform a new task in just a few trials and to perform that task more or less automatically after a few hundred" (p. 83, op.cit.). One could argue that this applies to tasks that stay constant and can in principle become largely automatic, but not necessarily to adaptive tasks like the current n-back task. In a progressively more difficult adaptive task, the participant might be taking into use new and more advanced strategies as the memory load keeps increasing, shifting for instance from a simple clustering strategy to more advanced grouping and comparison. However, in the current experiments, strategies were surprisingly stable after the initial phase with less than 20% strategy change after block 3 in Experiment 1 and less than 10% strategy change after block 4 in Experiment 2. This suggests that participants could adapt, for the most part, the same strategy to the increasingly demanding task. On Chein and Schneider's (2012) account, increasing performance once strategies have been fixed would be due to enhancement in the controlled execution or automatization processes that use the established strategy. Further studies are needed to determine to what extent these processes account for performance increase in an adaptive task after the initial stage. To take another example of learning that does not appear to fit to the three-stage timeline of skill learning, recent research has revealed instances where automaticity of performance is attained merely based on instructions without the need for overt practice (e.g., Cole et al., 2017; Longman et al., 2019). While Chein and Schneider (2012) do note that we can alter complex learned skills such as reading by just a single instruction, the framework itself does not address such situations.

Regarding the predictors of strategy use and n-back performance, neither age, educational attainment, metacognitive awareness, use of internal or external memory aids in everyday situations, or familiarization with responding were related to n-back performance. There may be several reasons for these null findings. The current age cap around 50 years was possibly too low to detect age-related decline in WM performance. For example, Dobbs and Rule (1989) observed WM declines beginning at age 60. Our measure of educational attainment was possibly affected by a limited number of participants at certain attainment levels. Additionally, some participants were probably still in the middle of their education, which could have confounded the results, also considering that WM has been shown to predict learning outcomes over time (Alloway & Alloway, 2010). Our null findings related to the MAI and MAQ exemplify the frequently observed null or small associations between self-report measures and objective cognitive task performance (e.g., Crumley et al., 2014; Duckworth & Kern, 2011), which could reflect issues with the scales or challenges in accurately reporting metacognitive ability or use of memory aids (for a discussion, see Harrison & Vallin, 2017). In the current context, questionnaires that are more specific and relatable to WM updating might yield statistically significant associations. Our measure of task routinization was, as discussed above, most likely misinterpreted by several participants and therefore of questionable value. Nevertheless, we see this line of investigation as highly relevant, and a possible fruitful extension would be to explore whether reaction times on easy task blocks interspersed along the task sequence could be used as an objective measure of automatization. On the other hand, two of our predictors were significantly related to n-back performance: level of detail in the open-ended strategy reports as well as verbal productivity (the level of detail being stronger of the two). As their predictive values concerning n-back performance suggested some overlap, strategy reports may at least partly reflect some non-strategy-related aspects such as verbal skills/writing ability (McCutchen, 1996; McNamara & Scott, 2001) and/or motivation.

Our second experiment showed that the list-based strategy queries, when placed after each block of the n-back task, facilitated participants' strategy use and task performance, which also supports the view that strategy use has a causal effect on performance (see Forsberg et al., 2020; Laine et al., 2018). We therefore caution against placing list-based strategy queries in the middle of a task, a test battery, or a longitudinal study where a listed strategy could be picked up by the participant and implemented in subsequent parts of the task/experiment. This result contradicts Dunlosky and Kane (2007), who report no reactive effects of repeated list-based strategy reports in a complex span task. We can only speculate why the results are contradictory, but one possible reason lies in the task paradigm. The n-back paradigm may come with a higher degree of novelty and feel more complicated than complex span that requires serial recall. This could lead participants to pick up strategies more often from a ready-made list in the n-back (and benefit from them) than in the complex span. Two interesting follow-up questions should be addressed in the future: (1) do even open-ended queries influence participants by prompting strategy-related metacognitive thinking; and (2) do open-ended questions related to strategy use (or questions about how a task was solved) in the middle of a test battery influence performance on subsequent tasks in a test battery through possibly enhanced metacognitive processing?

4.1. Limitations

Our study shows that evaluating strategy use is not an easy and clear-cut endeavor, and unsurprisingly, these challenges pose the main limitations of this study. As mentioned above, list-based strategy queries (at least when used repetitively, as was done here) can be problematic due to their potential to influence participants. Furthermore, we cannot be sure that a reported strategy is the only strategy that a participant has been using, or that participants are accurate in evaluating and reporting their thought processes (which is true for any kind of introspective account of strategy use). A specific challenge related to open-ended strategy reports is the scoring process that can introduce rater-related subjective bias, making it important to determine inter-rater reliability as was done here.[7] Another challenge stems from the fact that several participants gave only very brief or vague responses (e.g., reports of using memory or concentration), which explains the high number of participants that have been classified as using an "Other" strategy. Furthermore, some participants never answered the open-ended queries. These two strategy categories (No strategy and Other) also seem to represent the main sources of discrepancy between the open-ended and list-based responses (see Table 6). Hence, this discrepancy could possibly reflect challenges in introspection or in verbalizing internal thought processes, or possibly a lack of motivation in writing detailed descriptions. Non-responding, and possibly vague responding to some degree, could potentially be reduced by requiring some response or by running the experiment in a laboratory-based setting where semi-structured interviews could be used. Objective measures of strategy use (e.g., by analyzing reaction times, see Wu et al., 2008) could provide a solution to the problems related to introspection, but they undoubtedly face challenges of their own. At the same time, one should emphasize that the discrepancies were mostly limited to "No" and "Other" categories, whereas another strategy category (Grouping/updating) showed quite high convergence. Moreover, both strategy categorization

[7] Following the suggestion of one of the reviewers, we replicated our analyses after excluding the participants who did not use a strategy. This replication was done to ascertain that our results were not driven/confounded by those participants who never reported using a strategy during the n-back task. The results of these analyses are presented in Appendix D. Overall, the result patterns are very similar in both sets of samples. Note, however, that verbal productivity was not anymore a significant predictor of n-back performance, but this could be related to the smaller sample size and lower statistical power.

methods yielded significant associations between strategy type and n-back performance, which indicates that strategies play an important role in n-back performance.[8]

A potential limitation to the generalizability of the current findings is the adaptive task design that we employed. The adaptive nature of the task could result in more frequent changes in strategies due to changing and rising task demands. A static task (e.g., repeating 3-back sequences) might therefore show even faster stabilization of strategy use. Nevertheless, based on the present results, there is no reason to doubt that strategies are adopted very quickly in the n-back task, be it adaptive or not. A second possible limitation to the generalizability of the present results is presented by the multiple exclusion criteria that were implemented, which resulted in the exclusion of 34% (Experiment 1) and 26% (Experiment 2) of participants. Hence, it is possible that these results reflect cognitive performance of particularly healthy individuals rather than of a general population. A different potential limitation to the current study is presented by its online nature. This could increase error variance in WM task performance due to, e.g., differing testing environments. However, previous research has shown that online cognitive testing replicates cognitive processing effects observed in the lab (e.g., Enochson & Culbertson, 2015; Germine et al., 2012; Waris et al., 2017), which supports the quality of this data collection method.

The fact that the present evidence for an association between advanced strategy use (such as Grouping and comparison) and higher n-back performance stems from a small portion of the participants could also be seen as a limitation. However, one should point out that the pattern was similar in both experiments. Also our previous studies that have examined strategy use in n-back tasks as a whole indicate that only a limited number of participants come up with strategies that appear to be the most effective ones (Fellman et al., 2020; Forsberg et al., 2020; Laine et al., 2018). In line with this, strategy studies using simple and complex span tasks have also found that the majority of participants use less effective strategies (e.g., Bailey et al., 2011; Dunlosky & Kane, 2007). On a positive note, this can open a way to boost task performance by instructing non-strategic participants to employ an effective mnemonic technique. However, recent n-back training results indicate that an externally provided strategy leads only to short-term performance gains when compared with uninstructed n-back training (Fellman et al., 2020).

4.2. Implications and conclusion

The present results show fast adoption of strategy use in the majority of participants performing an unfamiliar WM task, and that the kind of strategy one adopts correlates with objective task performance. These results highlight dynamic within-task evolvement that remains invisible when only the summative scores are registered. Using a detailed timeline analysis, one could reveal a more heavy (but fleeting) task-initial engagement of executive functions, indicating that performance indices from the beginning vs. towards the end of task are measuring partly different constellations of cognitive processes. Such analyses can provide valuable insights into different complex, unfamiliar cognitive tasks, not only WM tasks. Skill learning could be a reason why cognitive tasks often show poor convergent validity: when the cognitive system adapts to each specific task through creating task-specific skills (including task-specific strategies), performance on that task is optimized but inter-task correlations are weakened. If this is the case, psychometric research should focus not only on average performance level, but also on the dynamics of how a task is learnt.

The role of strategy choice in WM performance raises an issue concerning WM capacity and its measurement. We concur with Simmering and Perone (2013) who argued that what we call "capacity" is not capacity in the classical sense (number of slots in a memory store), but an end product that emerges from multiple cognitive systems that are operative during task performance. In other words, one cannot tease apart storage and processing components, and one should not consider memory storage a constant individual feature across tasks. If WM capacity is an emergent property of all cognitive processes that are involved in task performance, discussions on the possible direction of causal relationships between WM capacity and strategy use (see Laine et al., 2018; McNamara & Scott, 2001) would be rendered obsolete as strategies are already embedded in capacity.

In conclusion, our study sheds light on the dynamics of WM task performance by showing that strategies are implemented very early during a WM task, and that these strategies are related to objective task performance. Future research should investigate how and why specific strategies are adopted and try to develop objective measures of strategy use.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2020.103211.

Author statement

Otto Waris: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration.

Jussi Jylkkä: Conceptualization, Methodology, Writing - Review & Editing.

Daniel Fellman: Conceptualization, Methodology, Formal analysis, Writing - Review & Editing.

Matti Laine: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Funding

This study was financially supported by the Academy of Finland (grants No. 260276 and 323251) and the Åbo Akademi University Endowment (grant to the BrainTrain project). OW was partially funded by the Academy of Finland Flagship Programme (grant No. 320162). DF received grants from the Signe and Ane Gyllenberg Foundation.

Open practices statement

Both Experiment 1 and 2 have been preregistered, and the data is available as electronic supplementary material.

Declaration of competing interest

None.

Acknowledgements

The research was supported/partially supported by the INVEST Research Flagship.

References

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

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge: Cambridge UP.

Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences, 6*(1), 47–52. https://doi.org/10.1016/S1364-6613(00)01819-2.

Baddeley, A. (2010). Working memory. *Current Biology, 20*(4), R136–R140. https://doi.org/10.1016/j.cub.2009.12.014

[8] Intriguingly, the effect of strategy type on n-back performance for the list-based reports was larger in Experiment 2 than in Experiment 1, which could depend on the potentially less ambiguous strategy questionnaire used in Experiment 2, but it could also be a chance finding.

Bailey, H., Dunlosky, J., & Hertzog, C. (2009). Does differential strategy use account for age-related deficits in working-memory performance? *Psychology and Aging, 24*(1), 82–92. https://doi.org/10.1037/a0014078.

Bailey, H., Dunlosky, J., & Kane, M. J. (2008). Why does working memory span predict complex cognition? Testing the strategy affordance hypothesis. *Memory & Cognition, 36*, 1383–1390. https://doi.org/10.3758/mc.36.8.1383.

Bailey, H., Dunlosky, J., & Kane, M. J. (2011). Contribution of strategy use to performance on complex and simple span tasks. *Memory & Cognition, 39*, 447–461. https://doi.org/10.3758/s13421-010-0034-3.

Bailey, H. R., Dunlosky, J., & Hertzog, C. (2014). Does strategy training reduce age-related deficits in working memory? *Gerontology, 60*, 346–356. https://doi.org/10.1159/000356699.

Borella, E., Carretti, B., Sciore, R., Capotosto, E., Taconnat, L., Cornoldi, C., & De Beni, R. (2017). Training working memory in older adults: Is there an advantage of using strategies? *Psychology and Aging, 32*, 178–191. https://doi.org/10.1037/pag0000155.

Brown, A. L. & Palincsar, A. S. (1982). *Inducing strategic learning from texts by means of informed, self-control training.* Center for the Study of Reading, Technical Report No. 262. University of Illinois at Urbana-Champaign, Center for the Study of Reading, Champaign IL.

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist, 26*(4), 543–570. https://doi.org/10.1080/13854046.2012.680913.

Carretti, B., Borella, E., & De Beni, R. (2007). Does strategic memory training improve the working memory performance of younger and older adults? *Experimental Psychology, 54*(4), 311–320. https://doi.org/10.1027/1618-3169.54.4.311.

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.

Chouliara, N., & Lincoln, N. B. (2015). Developing a questionnaire to assess the outcome of memory rehabilitation for people with neurological disabilities. *International Journal of Therapy and Rehabilitation, 22*, 470–477. https://doi.org/10.12968/ijtr.2015.22.10.470.

Cole, M. W., Braver, T. S., & Meiran, N. (2017). The task novelty paradox: Flexible control of inflexible neural pathways during rapid instructed task learning. *Neuroscience and Behavioral Reviews, 81*, 4–15. https://doi.org/10.1016/j.neubiorev.2017.02.009.

Crumley, J. J., Stetler, C. A., & Horhota, M. (2014). Examining the relationship between subjective and objective memory performance in older adults: A meta-analysis. *Psychology and Aging, 29*, 250–263. https://doi.org/10.1037/a0035908.

Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron, 70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018.

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging, 4*, 500–503. https://doi.org/10.1037/0882-7974.4.4.500.

Duckworth, A. L., & Kern, M. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality, 45*, 259–268. https://doi.org/10.1016/j.jrp.2011.02.004.

Dunlosky, J., & Kane, M. J. (2007). The contributions of strategy use to working memory span: A comparison of strategy assessment methods. *Quarterly Journal of Experimental Psychology, 60*, 1227–1245. https://doi.org/10.1080/17470210600926075.

Engle, R. W., Nations, J. K., & Cantor, J. (1990). Is "working memory capacity" just another name for word knowledge? *Journal of Educational Psychology, 82*(4), 799–804. https://doi.org/10.1037/0022-0663.82.4.799.

Enochson, K., & Culbertson, J. (2015). Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PLoS One, 10*(3), Article e0116946. https://doi.org/10.1371/journal.pone.0116946.

Fellman, D., Jylkkä, J., Waris, O., Soveri, A., Ritakallio, L., Haga, S., … Laine, M. (2020). The role of strategy use in working memory training outcomes. *Journal of Memory and Language, 110*. https://doi.org/10.1016/j.jml.2019.104064.

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.

Friedman, N. P., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory and Language, 51*, 136–158. https://doi.org/10.1016/j.jml.2004.03.008

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex, 86*, 186–204. https://doi.org/10.1016/J.CORTEX.2016.04.023.

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.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*, 847–857. https://doi.org/10.3758/s13423-012-0296-9.

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 1*, 103–111. https://doi.org/10.1016/j.dadm.2014.11.003

Harrison, G. M., & Vallin, L. M. (2017). Evaluating the metacognitive awareness inventory using empirical factor-structure evidence. *Metacognition and Learning, 13*, 15–38. https://doi.org/10.1007/s11409-017-9176-z.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron, 105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002.

Kaakinen, J. K., & Hyönä, J. (2007). Strategy use in the reading span test: An analysis of eye movements and reported encoding strategies. *Memory, 15*, 634–646. https://doi.org/10.1080/09658210701457096.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology, 55*, 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*, 4045. https://doi.org/10.1038/s41598-018-22396-5.

Longman, C. S., Liefooghe, B., & Verbruggen, F. (2019). How does the (re)presentation of instructions influence their implementation? *Journal of Cognition, 2*(1). https://doi.org/10.5334/joc.63.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8*, 299–325. https://doi.org/10.1007/BF01464076.

McNamara, D., & Scott, J. L. (2001). Working memory capacity and strategy use. *Memory & Cognition, 29*(1), 10–17. https://doi.org/10.3758/bf03195736.

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.

Morrison, A. B., Rosenbaum, G. M., Fair, D., & Chein, J. M. (2016). Variation in strategy use across measures of verbal working memory. *Memory & Cognition, 44*(6), 922–936. https://doi.org/10.3758/s13421-016-0608-9.

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.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*, 460–475. https://doi.org/10.1006/ceps.1994.1033.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*, 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

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., Lehtonen, M., Karlsson, L. C., Lukasik, K., Antfolk, A., & Laine, M. (2018). Test-retest reliability of five frequently used executive tasks in healthy adults. *Applied Neuropsychology: Adult, 25*, 155–165. https://doi.org/10.1080/23279095.2016.1263795.

Szmalec, A., Verbruggen, F., Vandierendonck, A., & Kemps, E. (2011). Control of interference during working memory updating. *Journal of Experimental Psychology: Human Perception and Performance, 37*(1), 137–151. https://doi.org/10.1037/a0020365.

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

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*. https://doi.org/10.3389/fpsyg.2017.01062.

Wu, S. S., Meyer, M. L., Maeda, U., Salimpoor, V., Tomiyama, S., Geary, D. C., & Menon, V. (2008). Standardized assessment of strategy use and working memory in early mental arithmetic performance. *Developmental Neuropsychology, 33*, 365–393. https://doi.org/10.1080/87565640801982.