Memory & Cognition (2020) 48:1484–1503
https://doi.org/10.3758/s13421-020-01066-w

Changes in Error Patterns during N-back Training Indicate Reliance on Subvocal Rehearsal

Weng-Tink Chooi [1] · Robert Logie [2]

Published online: 13 July 2020
© The Psychonomic Society, Inc. 2020

Abstract
Contemporary cognitive training literature suggests that training on an adaptive task produces improvements only in the trained task or near transfer effects. No study has yet systematically explained the mechanism behind improved performance on the N-back. In this study, we first investigated how improvements in an N-back task using eight pairs of phonologically similar words as stimuli occurred by examining error distributions of the task over training sessions. Nineteen participants (non-native English speakers) trained for 20 sessions over 5 weeks. We observed a reduction in false alarms to non-target words and fewer missed target words. Though the absolute number of phonological-based errors reduced as training progressed, the proportion of this error type did not decrease over time suggesting participants increasingly relied on subvocal rehearsal in completing the N-back. In the second experiment, we evaluated if improvements developed during N-back training transferred to tasks that relied on serial order memory using simple span tasks (letter span with phonologically distinct letters, letter span with phonologically similar letters, digit span forward, and digit span backward). Twenty-nine participants trained on the N-back and 16 trained on the Operation Span (OSPAN) for 15 sessions over 4 weeks. Neither group of participants showed improvements on any of the simple span tasks. In the third experiment, 20 participants (16 native English speakers) trained on the N-back for 15 sessions over 4 weeks also showed increasing reliance on subvocal rehearsal as they progressed through training. Self-report strategy use did not predict improvements on the N-back.

Keywords Cognitive training · Strategy development · OSPAN · N-back · Serial order recall

Introduction

Cognitive training studies have gained popularity in the past 15 years with some studies claiming that it can reduce attention deficit and hyperactivity disorder (ADHD) symptoms in children (Klingberg, Forssberg, & Westerberg, 2002) and improve general intelligence (S. M. Jaeggi, Buschkuehl, Jonides, & Perrig, 2008). The excitement surrounding cognitive training stems from findings that suggest transfer of improvements from the trained task to other untrained tasks. Near transfer occurs when training on a task improves not only the trained task but also a different task that measures the same construct, for example working memory. Both N-back and running back tasks (Bunting, Cowan, & Saults, 2006; Jaeggi, Buschkuehl, Perrig, & Meier, 2010) are commonly used measures of working-memory updating. If one trains on the N-back task, near transfer occurs when performance on the running back task improved after compared to before training or without training. Far transfer occurs when training on a task improves performance not only on the trained task but also on tasks from a different cognitive domain; for example, training on working memory might improve not only working memory but also tests of executive functions and of general intelligence. A decade's worth of literature has claimed that there are potential benefits of cognitive training in clinical populations such as patients with schizophrenia (e.g., Contreras, Tan, Lee, Castle, & Rossell, 2018; Genevsky, Garrett, Alexander, & Vinogradov, 2010; Subramaniam et al., 2018), chronic brain traumatic injuries (e.g., Han, Chapman, & Krawczyk, 2018), stroke (e.g., De Luca et al., 2018; Westerberg et al., 2007), multiple sclerosis (e.g., Dardiotis et al., 2018; Vogt et al., 2009), chronic fatigue syndrome (e.g., McBride et al., 2017), older adults with mild cognitive impairments (e.g.,

✉ Weng-Tink Chooi
wengtink@usm.my

[1] School of Social Sciences, Universiti Sains Malaysia, Gelugor 11800, Malaysia

[2] Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK

Bottiroli, Cavallini, & Vecchi, 2008; Liu et al., 2016), as well as people living with HIV (e.g., Towe, Patel, & Meade, 2017). However, mechanisms underlying reported effects of cognitive training, specifically working-memory training, are not fully understood, and whether such effects are reliable and replicable is controversial (Au, Buschkuehl, Duncan, & Jaeggi, 2016; Chooi & Thompson, 2012; De Simoni & von Bastian, 2018; Gathercole, Dunning, Holmes, & Norris, 2019; Melby-Lervåg, Redick, & Hulme, 2016; Redick et al., 2013; Shipstead, Hicks, & Engle, 2012; Simons et al., 2016). Here, we focus on providing insight into the changes during training with implications for any subsequent transfer. That is, instead of a focus on the possible transfer effects, our aim was to enhance the theoretical understanding of what happens to participants' cognition during training, and whether that understanding may be important for explaining why transfer effects do or do not occur.

The first meta-analysis on working-memory training (Melby-Lervåg & Hulme, 2013) summarized that working-memory training programs produced short-term specific training effects that do not generalize to untrained nonverbal and verbal abilities, inhibitory processes in attention, word decoding, or arithmetic. Based on results from the most robust designs incorporating randomized trials and comparison of an intervention group with both passive and active control groups, this meta-analysis showed that there is no evidence of transfer effects from working-memory training to measures of nonverbal ability. A more recent multi-level meta-analysis focusing only on N-back training reported a medium effect size for task specific improvements and small effect sizes for transfer to other working memory tasks, cognitive control tasks, and general intelligence (Soveri, Antfolk, Karlsson, Salo, & Laine, 2017). The biggest unresolved issue of working-memory training programs is that they "*do not appear to rest on any detailed task analysis or theoretical account of the mechanisms by which such adaptive training regimes would be expected to improve working-memory capacity*" (Melby-Lervag & Hulme, 2013). Logie (2012) suggested several possible explanations for any observed changes that occur during or after training. The simplest account is the well-established observation that people get better at a task when they practice that task because either some aspects of task performance become automated or people develop strategies during training that allow them to perform the task more efficiently. Any benefit of the training may then transfer to performance of other tasks because either the automated skills or the newly developed strategies are also helpful in performing these other tasks.

The investigation of strategy development after prolonged and intensive practice on a task and possible transfer to other tasks has a long history (e.g., Donchin, Fabiani, & Sanders, 1989; Thorndike & Woodworth, 1901; see reviews in Colley & Beech, 1989). Thorndike and Woodworth (1901) suggested in their *identical elements* theory that transfer could happen between tasks if both tasks shared common elements in executing the tasks. This theory is implied as the basis of theoretical rationale for most training studies in the early 2000s (e.g., Jaeggi et al., 2008), though not directly cited in these recent publications. In contrast to that line of reasoning, more recently Laine, Fellman, Waris, and Nyman (2018) have shown that, for some cognitive skills, and with the right procedures, extensive training might not be necessary. These researchers produced transfer after just one training session, in stark contrast with a typical 4- to 6-week training study of 10–20 training sessions in many previous studies. Participants in the Laine et al. (2018) study were allocated to three groups, with two of the groups trained on the digit N-back task, and the third group with no training served as passive control. Participants in one of the training groups received instructions on how to complete the N-back (the other training group served as active control), and their rates of improvement were significantly higher than for those who trained without receiving instructions on how to complete the task. Improvements gained from training with instructions transferred to untrained N-back tasks using letter and color stimuli, digit span, selective updating of digits, and running memory task using digits. These improvements were significantly higher than improvements seen in the active (training without instructions) and passive control groups (Laine et al., 2018).

Some of the recent literature (e.g., Bailey, Dunlosky, & Hertzog, 2014; Dunning & Holmes, 2014; Peng & Fuchs, 2015) suggests that strategies are in play but it is not clear what these strategies might be, and what exactly changes during and after training that causes improvements on the trained task. One possibility for investigating this in detail would be to examine the errors that participants make, and whether the types of errors that they generate change during training (e.g., Logie, Baddeley, Mane, Donchin, & Sheptak, 1989). In Experiment 1 of our study, we investigated how errors that participants made changed during training on an adaptive verbal N-back task. In Experiment 2, we investigated whether adaptive training on the N-back could improve performance on simple span tasks, and whether changes in error patterns observed in Experiment 1 could allow us to predict whether or how performance on these other tasks might benefit from changes that occurred during training. In Experiment 3, we investigated how self-reported strategy development, general cognitive ability, and working-memory capacity influenced task performance in the N-back task.

Experiment 1

The main objective of this experiment was to identify types of errors made when participants trained on the N-back task and

evaluate how these errors changed during training. We did not include a control group as identifying transfer effects was not a goal in this experiment. We designed an adaptive verbal N-back task using eight pairs of homophones to examine the influence of phonological coding in executing the task. We hypothesized that there should be more errors (false alarms) generated by words sounding similar to target words at the beginning of training. As participants became aware of such interference, they would develop strategies to reduce interference caused by lure words, and we would see a reduction in this type of error towards the end of training. As the objective of this experiment was to evaluate changes in error patterns during training, we did not administer any assessments pre- and post-training to document any transfer effects.

Methods

Participants and training sessions

Participants were recruited via email advertisement and word of mouth among university students in Malaysia. Interested participants were asked to email the researcher for more information. Study information and a digital copy of the consent form were emailed to prospective participants. If they agreed to participate, they arranged to meet with a research assistant in the study. All participants signed an informed consent form at the start of the first session. They trained for 20 sessions over 5 weeks, and each training session lasted about 30 min. Participants were compensated Malaysian ringgit (RM)5 for each training session. They were rewarded a RM20 bonus if they completed all training sessions for a total of RM120 (equivalent to about £22 or US$30). Ethics approval was obtained from the Human Research Ethics Committee, Universiti Sains Malaysia.

Thirty university students (24 females) participated in the study (mean age: 22.6 years; age range: 20–35 years). All were non-native English speakers, but they all had completed at least 11 years of learning English as second language and passed the English (written) paper/subject in their secondary school exit examination. Nineteen participants completed all 20 training sessions. Six participants missed three training sessions consecutively and their participation was terminated. Five participants withdrew from the study due to technical and scheduling issues.

The N-back task

An adaptive verbal N-back task was created using eight pairs of homophone words – ate-eight, blue-blew, feel-fill, hear-here, made-maid, pool-pull, soul-sole, and vain-vein. Each word was displayed on the screen for 2,000 ms with an inter-stimulus interval of 500 ms. There were 40+n words with 20% target words and 10% lures in one block. Participants had to respond to each word stimulus whether it was the same as (match) or different from (no match) the word that was 'n' back from the current word. A target was a stimulus that met the N-back criteria for a match. Stimuli that were target words but presented at position n-1 (closer to the currently presented word than the target n) or n+1 (further back than the target n) were called *lure position* words. All other stimuli presented within a block were non-target words or *foils* (see Fig. 1). The value of 'n' in the N-back task was referred to as load.

The task was programmed to adapt its difficulty level or load to each participant's ability by automatically increasing the load (value of n) when the participant had attained an accuracy score of more than 90% correct target detection. If the score fell below 60% accuracy, the load was reduced. The participant continued performing the task with the same load if the accuracy score stayed between 60% and 89%. Participants started at load 2 (2-back) in each training session, and the task stopped after the number of training blocks specified at the beginning of training session was achieved (10 or 15 blocks).

Data analysis

We analyzed our results using repeated-measures ANOVA on SPSS and Bayesian repeated-measures ANOVA on JASP. We aggregated data from 20 sessions into four training phases, with data from five training sessions in each training phase. We selected data from loads 2, 3, and 4 only as all participants attempted load 4 at least once over the entire study. We averaged total errors, total targets missed, total false alarms due to non-target (foil) words and lures separately for each load in each training phase. Targets missed were "no match" responses to target words and false alarms were "match" responses to foils (non-target words). The dependent variables in the repeated-measures ANOVA analyses were total errors, targets missed, foil (non-target word) false alarms, lure word false alarms, lure position n-1 false alarms, and lure

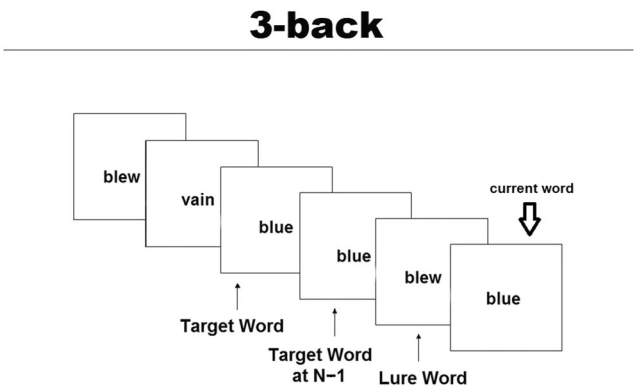

Fig. 1 N-back task illustrating when a stimulus is identified as a target

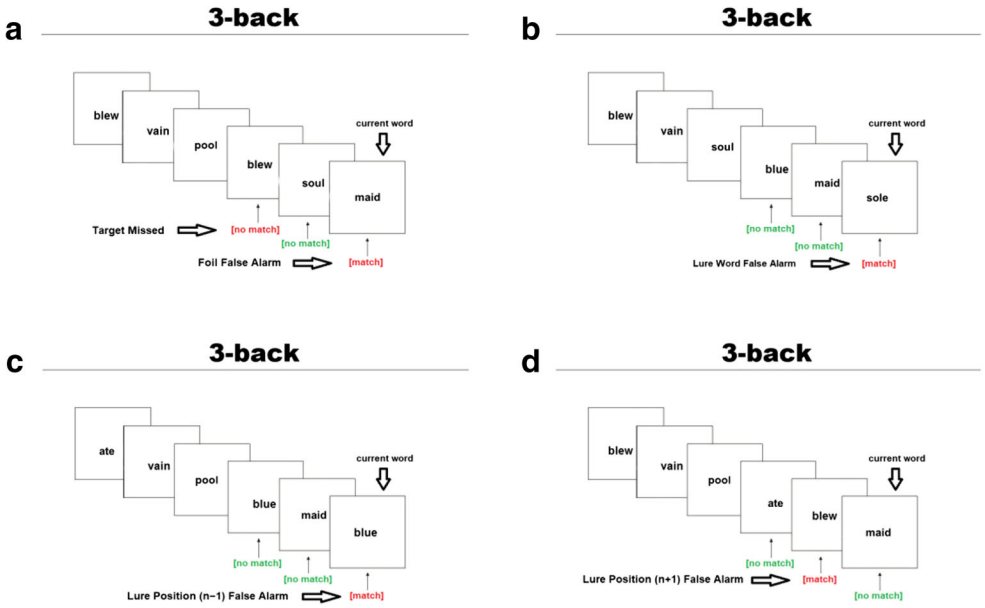

Fig. 2 Types of errors: targets missed, foil false alarms, lure word false alarms, lure position (n-1) / (n+1) false alarms. Responses from participants ("match" or "no match") in red font are incorrect responses while responses in green font are correct responses

position n+1 false alarms over four training phases (see Fig. 2 for types of errors).

Types of errors

Lists of 40+n words were randomly generated in every training session. The training task was programmed to generate 10% of lures from total words in each list. "Lure word false alarms" were errors due to homophone of the target word in n-back position. "Lure position n-1 / n+1 false alarms" were errors caused by a word identical to the target word but at a surrounding target position (one position either before or after the target position). Lure positions were included to avoid responses based on familiarity (to ensure participants stay focused throughout training).

In Fig. 2, a block of six words in the 3-back condition was presented to illustrate each error type. In panel A, if a participant responded "no match" to the word *blew*, it was coded as "target missed" because *blew* appeared three words before. If a participant responded "match" to the word *maid*, it was coded as a "foil false alarm" as the target word three positions back was *pool*. In panel B, a participant made the correct response "no match" to *blew* because the target word was the homophone *blue*. An incorrect response of "match" to the word *sole* would be coded as "lure word false alarm" as the target word 3-back was the homophone *soul*. In panel C, the incorrect response of "match" to the word *blue* was coded as "lure position (n-1) false alarm" because the target word was at 2-back. Similarly, in panel D, the incorrect response of "match" to the word *blew* was coded as "lure position (n+1) false alarm" as the target word was at 4-back.

Results

Twenty-one participants who trained for at least 18 sessions showed improvements on the N-back task at the end of training and were included in the current analyses. There was a significant increase of mean load attempted from session 1 (mean = 2.31, SD = 0.27) to session 20 (mean = 4.98, SD = 1.27; $t(18) = 11.54$, $p < 0.001$; $BF_{10} = 1.11 \times 10^7$; see Fig. 3).

Data from loads 2, 3, and 4 were included in an overall analysis because every participant completed 4-back at least once by session 18. Overall, total errors including targets missed and total false alarms decreased over time (Fig. 4). Mean errors of each type of false alarm reduced over time (Fig. 5, left panel). The change in proportion of each type of false alarm (foil, lure word, lure position n-1, and lure position n+1) over total false alarms showed that foil false alarms

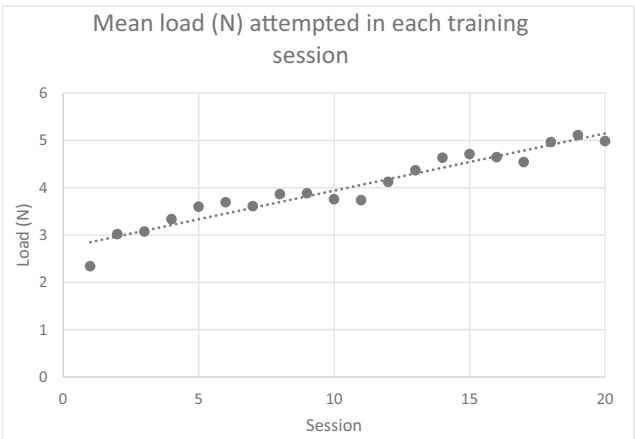

Fig. 3 Mean load (n) attempted in each training session across 20 sessions

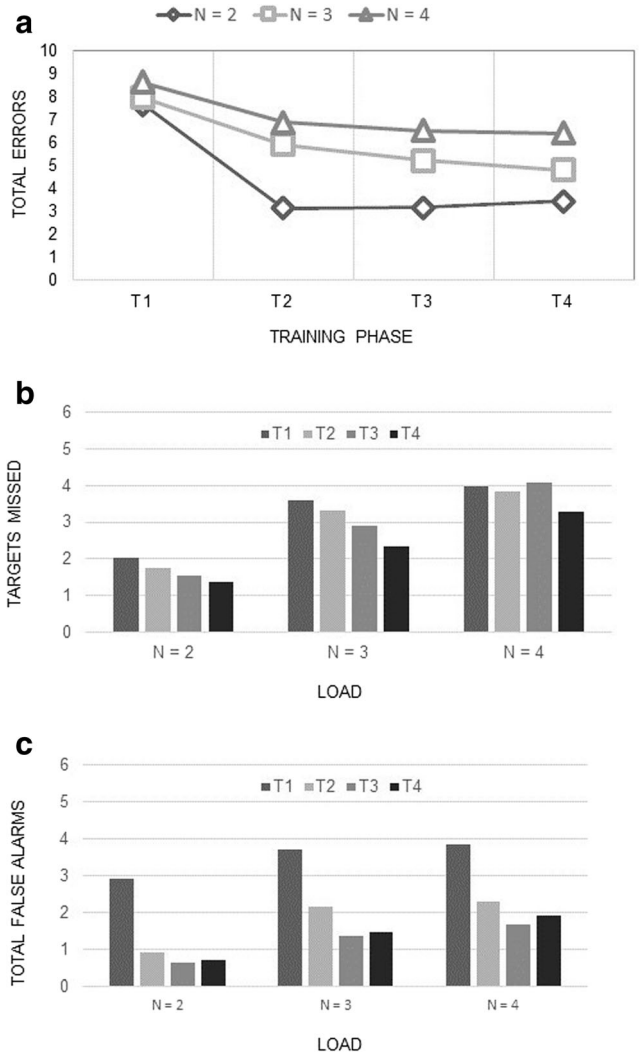

Fig. 4 Trends of total errors over four training phases (20 sessions) among participants in Experiment 1. (**A**) Total errors; (**B**) targets missed; (**C**) total false alarms

showed a more rapid reduction over time compared to lure word false alarms (Fig. 5, right panel). Results from ANOVA repeated measures and Bayesian repeated measures of the whole sample are summarized in Table 1.

Sensitivity index (d') and bias (c)

Using parameters from signal detection theory, sensitivity index (*d'*) and bias (*c*) were calculated for every participant and averaged over each training phase. *d'* is estimated from the difference between hit-rate and false-alarm rate:

$$d' = Z(hit\ rate) - Z(false\ alarm\ rate)$$

There was an increase of *d'* over time (see Fig. 6, left most column; $F(3) = 6.46$, $p < 0.001$; $BF_{10} = 2905.7$). A negative value of *c* indicated a higher tendency to respond "match" to stimuli, and a positive value of *c* indicated more responses of "no match" to stimuli.

Discussion

Our findings revealed that total errors decreased over time as training progressed, which was reflected in improved performance overall. The types of errors that contributed most to total errors made as training progressed were from targets missed and homophone lure word false alarms. The relatively infrequent presence of homophone lures (10%) allowed overall performance to improve when these lures were not a source of interference in a trial. At the same time, the homophones acted as a diagnostic for how participants were performing the task. The absolute number of lure word false alarms decreased with training in the 2-back and 3-back conditions, but the rate of decline was not as rapid as the other types of errors, resulting in an increase in the proportion of errors resulting from phonologically based (homophone) errors. This coupled with an overall reduction in the total number of errors suggested that participants increasingly relied on phonologically based coding while other strategies were used less frequently or were abandoned as training progressed. This observation was the opposite of our hypothesis that predicted fewer errors due to phonologically similar sounding words. The increased reliance on phonological coding suggested participants engaged in more sub-vocal rehearsal or depended more on an auditory memory trace.

The overall improvement in performance with training was also clear from the increase in sensitivity index, *d'*, over time. Moreover, participants made more false alarms at the beginning of training, and this was reflected by negative *c* values. A negative value of *c* indicated a more liberal response of responding incorrectly with "yes" or "match" and, hence, more false alarms. A positive value of *c* indicated a more conservative response of responding "no" or "no match" to stimuli, so when target item was present in the N-back position, more targets were missed.

Given that participants appeared to rely more on phonological coding and sub-vocal rehearsal as they trained on the N-back task, it may be that this practice gradually improved one's ability to remember phonologically encoded sequences in serial order after training. In Experiment 2, we included simple span tasks with phonologically similar and phonologically different items to test that hypothesis.

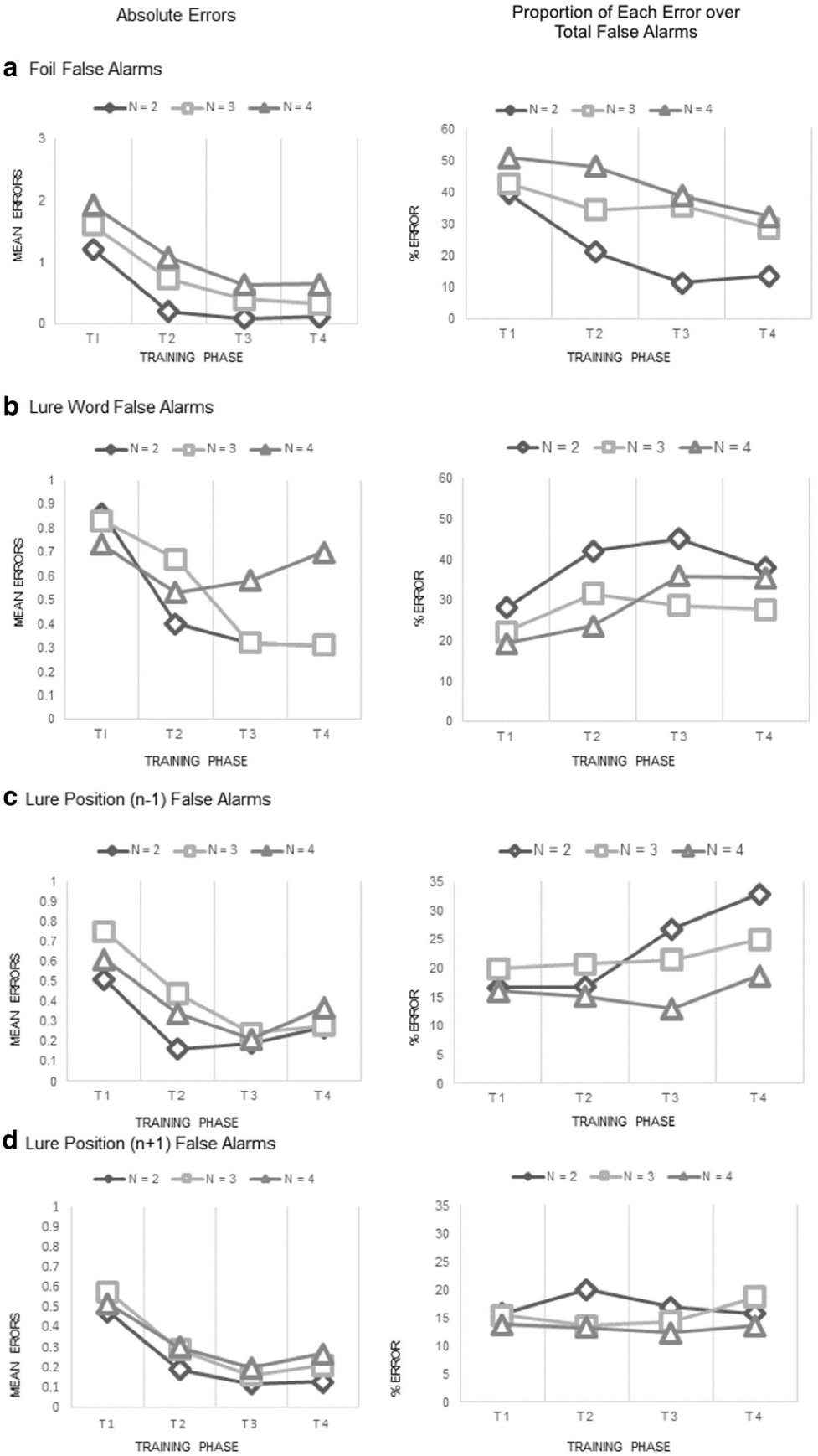

Fig. 5 Mean absolute errors and proportion of false alarms among Experiment 1 participants

Experiment 2

The results from Experiment 1 indicated that during training, participants increasingly relied on phonologically based coding of items, and possibly sub-vocal rehearsal, that led to an overall improvement in performance on the n-back task. This might suggest that N-back training allowed for practice in the use of a phonological-based rehearsal strategy to support task performance. Therefore, in Experiment 2, we attempted to replicate the error distribution observed in Experiment 1 and explored whether this form of N-back training might show transfer to improvements in four serial-order recall tasks, each with different verbal memoranda administered before and after training. This then was compared with the impact of training on a different working-memory task, known as operation span (OSPAN), acting as an active control group. We did not expect participants to improve on the effectiveness of sub-vocal rehearsal based on phonological codes in the OSPAN training group as the nature of complex span tasks that alternated between processing a task and remembering an item to be recalled later did not provide opportunities for participants readily to engage in sub-vocal rehearsal to remember the items to be recalled. Performance on the OSPAN required successful retrieval of items displaced from primary memory for immediate recall into secondary memory by the processing component of the task, and successful retrieval is thought to depend on a cue-dependent search process (Unsworth & Engle, 2006).

In Experiment 2, we compared performance between two training groups – the N-back and operation span (OSPAN) – on selected simple span tasks after 4 weeks of training. We predicted that N-back training would improve performance on serial-order tasks and OSPAN training would not have any effect on performance on such tasks. In this experiment, the OSPAN training group served as an active control group.

Methods

Participants

Participants were undergraduate students minoring in psychology and recruited via in-class announcements. Recruitment and informed consent procedures followed those of Experiment 1. All were non-native English speakers, but they all had completed at least 11 years of learning English as second language and passed the English (written) paper/subject in their secondary school exit examination. Participants trained for 15 sessions over 4 weeks, and each training session lasted about 30 min. Participants were compensated RM10 for the first and last training sessions and RM5 for each subsequent training session. They were rewarded a RM15 bonus if they completed all training sessions for a total of RM100 (equivalent to about £18 or US$25). Ethics approval was obtained from the Human Research Ethics Committee, Universiti Sains Malaysia.

Participants were assigned to one of two training paradigms – the N-back or the Operation Span (OSPAN). Twenty-nine students (mean age: 23.0 years; age range: 21–24 years; 27 females) trained on the N-back task and 26 students (mean age: 22.9 years; age range: 21–24 years; 23 females) trained on the OSPAN task. Both groups trained for approximately 30 min 15 times in 4 weeks. One student in the OSPAN group was excluded due to three consecutive absences after completing 12 training sessions. Compared to the first experiment, there was a higher retention rate in this experiment.

Table 1 Summary of repeated-measures ANOVA and Bayesian repeated-Measures ANOVA of total errors and types of errors generated during n-back training sessions

	Training phase main effect		Load main effect		Interaction Training Phase x Load	
	F-value ($F(3)$)	Bayesian Factor (BF)	*F*-value ($F(2)$)	Bayesian Factor (BF)	*F*-value ($F(6)$)	Bayesian Factor (BF)
Total errors	30.19**	3.73×10^{11}	26.91**	2.50×10^{6}	3.81*	3.95
Targets missed	4.17*	1.03	44.62**	5.60×10^{19}	1.77	0.13
Foil false alarms	46.16**	6.90×10^{22}	15.21**	94.48	1.11	0.08
Lure word false alarms	11.70**	167,825.88	16.78**	1278.73	2.25*	0.62
Lure position n-1 false alarms	13.06**	8.54×10^{7}	2.62	0.90	1.91	0.14
Lure position n+1 false alarms	23.88**	2.29×10^{13}	1.74	0.17	0.75	0.06

**$p < 0.001$

*$p < 0.05$

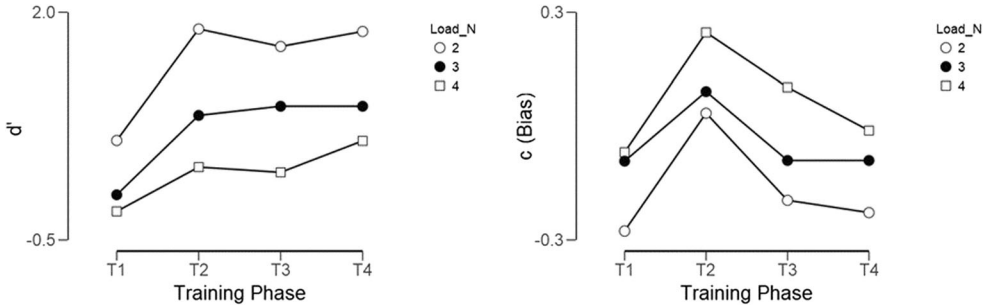

Fig. 6 Sensitivity index, *d*', across four training phases (**left panel**) and Bias, *c*, across four training phases (**right panel**) among Experiment 1 participants

Procedure

In the first session, participants gave consent to participate and they completed four criterion tasks consisting of phonologically similar sounding and distinct sounding letter spans, and forward and backward digit spans. After completing the criterion tasks, participants began their first training session. The first session lasted approximately 60 min. Each subsequent training session lasted approximately 30 min. During the last training session, session 15, participants first completed the training task before completing the criterion tasks in the same order.

Criterion tasks

Four computerized simple span tasks were designed as criterion tasks for the experiment. The tasks used as criterion tasks were as follows:

- Letter Span – Phonologically Similar: Similar sounding letters (B, C, D, E, G, P, T, V) were used as stimuli in this task. Each stimulus was presented one at a time at the center of the screen, and participants were asked to recall all the letters in the order presented when the instruction to recall appeared. The task started with three letters, and this automatically increased by one letter after participants successfully recalled all the letters presented in the correct order in two out of three trials. The task stopped when a participant failed to recall the letter sequence accurately on two of the three trials.
- Letter Span – Phonologically Distinct: F, H, J, L, M, Q, S, U were used as stimuli in this task. The task operated the same way as Letter Span – Similar.
- Digit Span – Forward: Digits 0–9 were used as stimuli in this task. The task operated the same way as the letter spans, with digits presented in a random order for each sequence.
- Digit Span – Backward: Digits 0–9 were used as stimuli in this task. This task operated in the same way as letter span and digit span forward except that participants had to recall the digits presented in the reverse order.

The Operation Span (OSPAN) training task

We designed an adaptive OSPAN task using the same eight pairs of homophone words in the N-back task as words to be remembered. Participants had to respond "true" or "false" to mathematical equations that appeared. After responding, a word appeared for 2,000 ms followed by the next equation to be evaluated. At the end of a series of equations each followed by a word, participants had to recall all the words in the order presented (see Fig. 7). Each training session started with three equations and three words. After accurately recalling all the words presented in the correct order and obtaining an accuracy above 90% in the math component, the task automatically increased the span by 1. If participants failed to recall the words presented in the right order twice or obtain an accuracy above 90%, the task reduced the span by 1. Participants completed about 35–40 trials in a 30-min session.

Data analysis

Data analysis was as for Experiment 1. We aggregated data from 15 sessions into three training phases, with data from five training sessions in each training phase. Only N-back training participants were included in the error analysis. We selected data from loads 2, 3, and 4 to replicate analyses from Experiment 1. All but one participant attempted load 4 at least once over the entire study.

Results

N-back participants showed improvements on the task after 15 sessions of training. There was a significant increase of mean load attempted from session 1 (mean = 2.47, SD = 0.38) to session 15 (mean = 5.03, SD = 1.54; t (28) = 9.75, p <0.001; BF_{10} = 6.08 x 10^7). OSPAN participants also showed improvements on the task after 15 sessions of training. There was a significant increase of mean load (or list length) attempted from session 1 (mean = 3.63, SD = 0.80) to session 15 (mean = 5.08, SD = 1.62; t (24) = 5.63, p <0.01; BF_{10} = 2488).

Data from loads 2, 3, and 4 were presented in the following analyses. Overall, total errors including targets missed and total false alarms decreased over time (Fig. 8).

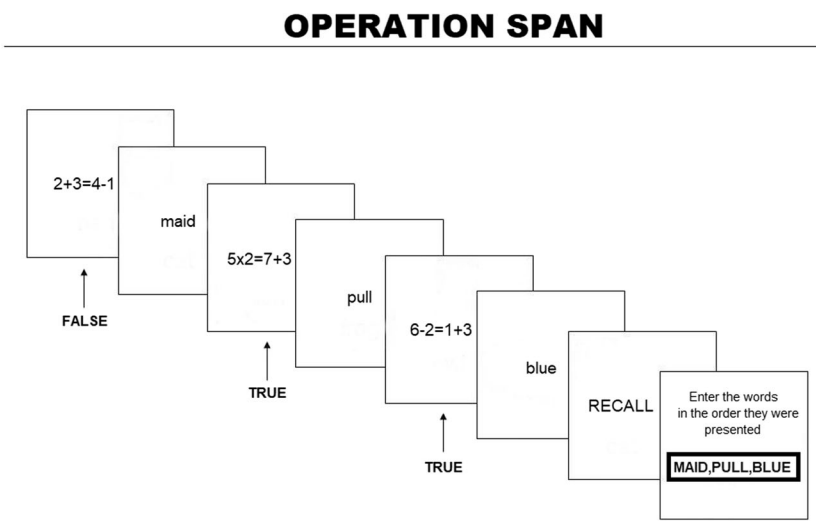

Fig. 7 Description of the Operation Span task

Results from ANOVA repeated measures and Bayesian repeated measures of the whole sample were summarized in Table 2. Mean errors of each type of false alarm reduced over time (Fig. 9, left column). The change in proportion of each type of false alarm (foil, lure word, lure position n-1, and lure position n+1) over total false alarms showed that foil false alarms decreased over time but false alarms due to lure word and lure position (n-1) did not decrease. False alarms due to lure position (n+1) showed a slight increase over time (Fig. 9, right column).

Sensitivity index (d') and bias (c)

There was a significant main effect of training phase ($F(2) = 22.09$, $p < 0.001$; $BF_{10} = 2.82 \times 10^{10}$) and load ($F(2) = 22.30$, $p < 0.001$; $BF_{10} = 1013.6$) in *d*' (see Fig. 10). There was a significant main effect of load ($F(2) = 38.91$, $p < 0.001$; $BF_{10} = 8.00 \times 10^{7}$) and interaction effect of training phase and load ($F(4) = 5.42$, $p = 0.001$; $BF_{10} = 33.2$) in *c*. Main effect of training phase was not significant ($F(2) = 1.63$, $p = 0.21$; $BF_{10} = 0.25$). The whole sample started with more liberal responses in the 2-back and more conservative in the 3- and 4-back conditions.

Criterion Tasks Analyses between training (N-back or OSPAN) groups

There were no significant differences between training groups at baseline and post-training for all the criterion tasks. There were also no significant improvements in any criterion task among N-back and OSPAN participants after training according to paired t-test analyses (see Table 3).

Discussion

In this experiment, we compared effects of two different training tasks – the N-back and OSPAN. We predicted that participants who trained on the N-back would improve on recalling serial order information in simple span tasks, but they did not. Those who trained on the OSPAN also did not show any significant changes on any of the criterion tasks after 4 weeks of training, which confirmed our prediction that repeatedly practising the OSPAN did not improve participants' ability to remember and recall items in serial order. The lack of a phonological similarity effect in the criterion tasks before and after training was perhaps due to some of the letters in the Letter Span Distinct sounding similar (e.g., F, S, H) to the non-native English speakers in the sample. These letters were chosen from letter sets that have been selected as phonologically distinct in previous studies (e.g., Kane et al., 2004).

In this experiment, we observed similar patterns of error changes as those in Experiment 1. All errors decreased over time but the rate of decrease for lure word errors was not as rapid as the other types of errors, resulting in an increase in phonologically based errors as training progressed. This again supports the conclusion that participants increasingly relied on phonological coding of stimuli as they improved performance on the adaptive N-back task.

Experiments 1 and 2 were conducted in Malaysia where participants were non-native English speakers. We conducted a third study in the United Kingdom with native English-speaking participants to investigate whether the results from Experiments 1 and 2 would replicate. We also asked participants what strategies they employed in performing the N-back task, and how often those strategies were adopted over the course of training using a self-report questionnaire.

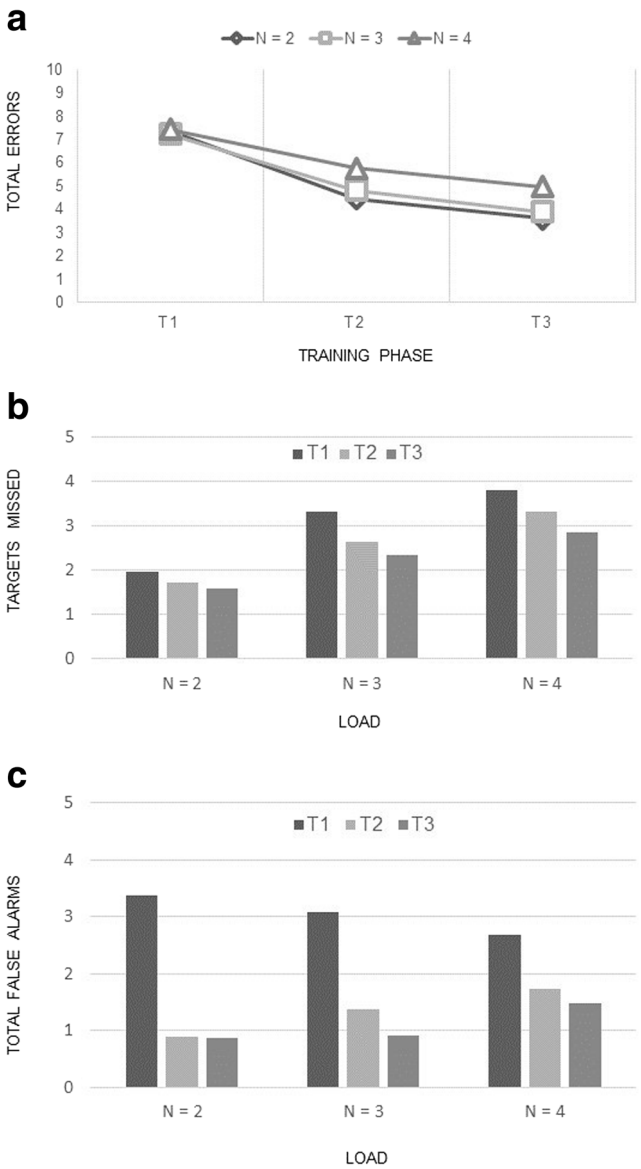

Fig. 8 Trends of total errors over three training phases (15 sessions) among participants in Experiment 2. (**A**) Total errors; (**B**) targets missed; (**C**) total false alarms

Experiment 3

It is well established that people improve on a task after repeated practice. Logie (2012) suggested that some aspects of a trained task become automated because people develop strategies during training that enable them to perform the task more efficiently. In Experiment 3, we provided a questionnaire that listed several possible strategies that could be used to perform the N-back. We investigated how self-report use of strategies could have affected gains in N-back training performance. The same N-back training paradigm in Experiment 1 was employed in this experiment. Our overall aim in these studies is to investigate changes during training, and our focus is on what changes during N-Back adaptive training. An active control group or OSPAN training, as used in Experiment 2, may involve different changes in cognition from those that occur during N-Back adaptive training. Moreover, there were no significant transfer of training effects in Experiment 2 for either group. Results from an active control group seemed unlikely to be informative for Experiment 3, and only a passive control group was included. We also included three additional criterion tasks in Experiment 3 – two working-memory capacity tasks (including OSPAN) and one reasoning ability test. We included the new tasks to evaluate if baseline working-memory capacity and general cognitive ability measured by a visual matrix reasoning ability test affected rate of improvements in N-back training. We did not predict any changes to these new criterion tasks as well as the others from Experiment 2.

Finally, in Experiments 1 and 2, although all participants were fluent in English, none were native speakers, and there was a suggestion that the manipulation of phonological similarity in the memoranda was not wholly effective, particularly in Experiment 2. Therefore, Experiment 3 was conducted in the UK, with a view to investigating whether results from Experiments 1 and 2 would replicate with native speakers of the English language.

Methods

Participants

There were two groups of participants – N-back training group and a passive control group. Participants in the training group completed two testing sessions and 15 training sessions spread across 3–4 weeks. They first completed a battery of cognitive tests consisting of working-memory capacity tests, simple span tasks including letter span tasks and digit span tasks from Experiment 2, and a matrix reasoning test before they started training on the N-back task (pre-test). After completing 15 training sessions, participants completed the same battery of cognitive tests (post-test). Participants in the control group completed the battery of cognitive tests on two separate occasions that were 3 weeks apart.

Participants were recruited via the online career search portal of the University of Edinburgh, word of mouth from participants in the study, and mailing lists in the department. Twenty-three participants initially signed up for the training study and 20 (mean age: 23.7 years; 17 females) completed the training sessions and pre- and post-training testing. Four who completed training were non-native speakers but had spent many years living in an English-speaking country/environment. Those who completed training and both testing sessions were compensated £90 (~US$116). There were 12 participants (mean age: 23.6 years; six females; one non-native English speaker) in the passive control group, and they received £30 (~US$40) after completing two testing sessions.

Table 2 Summary of repeated-measures ANOVA and Bayesian repeated-measures ANOVA of total errors and types of errors generated during n-back training sessions

	Training phase main effect		Load main effect		Interaction Training Phase x Load	
	F-value (*F*(2))	Bayesian Factor (BF)	*F*-value (*F*(2))	Bayesian Factor (BF)	*F*-value (*F*(4))	Bayesian Factor (BF)
Total errors	31.87**	3.38×10^{16}	17.65**	27.97	1.58	0.15
Targets missed	13.87**	2819.08	101.56**	6.27×10^{20}	0.77	0.06
Foil false alarms	27.55**	6.46×10^{15}	2.43	0.20	1.72	0.12
Lure word false alarms	27.54**	1.51×10^{6}	3.62*	0.96	4.40*	14.61
Lure position n-1 false alarms	29.25**	7.26×10^{9}	5.20*	2.06	2.74*	0.96
Lure position n+1 false alarms	27.65**	5.57×10^{10}	0.17	0.06	2.46	0.69

**$p < 0.001$

*$p < 0.05$

The battery of cognitive tests

Participants completed the following tests twice in the same order both times and took about 75–90 min to complete them. The tests were administered in the following order:

- Operation Span (OSPAN) shortened version (Foster et al., 2015): In this task, participants alternated between verifying a mathematical operation (processing component) and remembering the letter presented after the mathematical operation. Participants completed one block consisting of 25 mathematical problems and letters with trials varying in length (3–7).
- Symmetry Span shortened version (Foster et al., 2015): Participants alternated between judging whether figures presented were symmetrical or not and remembering the position of colored squares in a 4 x 4 matrix. Participants completed two blocks consisted of 28 symmetry judgments and trial sizes (number of colored square positions) between 2 and 5.
- Letter Span – Phonologically Distinct: as in Experiment 2
- Letter Span – Phonologically Similar: as in Experiment 2
- Digit Span Forward: as in Experiment 2
- Digit Span Backward: as in Experiment 2
- Raven's Advanced Progressive Matrices (RAPM) (Raven, Court, & Raven, 1977): There were 36 items in this test used as a measure of reasoning. Odd-numbered items were used at pre-test and even-numbered items were used at post-test. Each item presented figures or patterns in a 3 x 3 matrix with the bottom right corner blank. Participants had to decide which of the eight options presented below the matrix best fit or complete the matrix. Participants were given 10 min to complete as many items as possible in each session (pre- and post-test).

The training task – N-back

A verbal N-back task using eight pairs of homophone words was used as the training task for this study. See task description in Experiment 1.

Strategy questionnaire

We provided a questionnaire that listed several possible strategies (see Table 6) that could be used to perform the N-back after every five training sessions. Participants were asked to rate how often they used those strategies on a 6-point Likert scale (1 for never using the specified strategy and 6 for always using the specified strategy). They were also asked to report other strategies that they used if those were not listed in the questionnaire and rate how often they use those strategies.

Data analysis

Data analysis was as for Experiments 1 and 2. First, we analyzed N-back training data from loads 2, 3, and 4. We then compared baseline and post-training performance on reasoning ability, working-memory capacity as measured by complex span tasks and simple span tasks.

To analyze whether or how strategy development influenced performance improvements in N-back during training, reported strategies were categorized into characteristics such as rehearsal, chunking, visual imagery, and other. Strategy categories were then included as predictors in regression analyses with average load (N) attained at session 15 as the

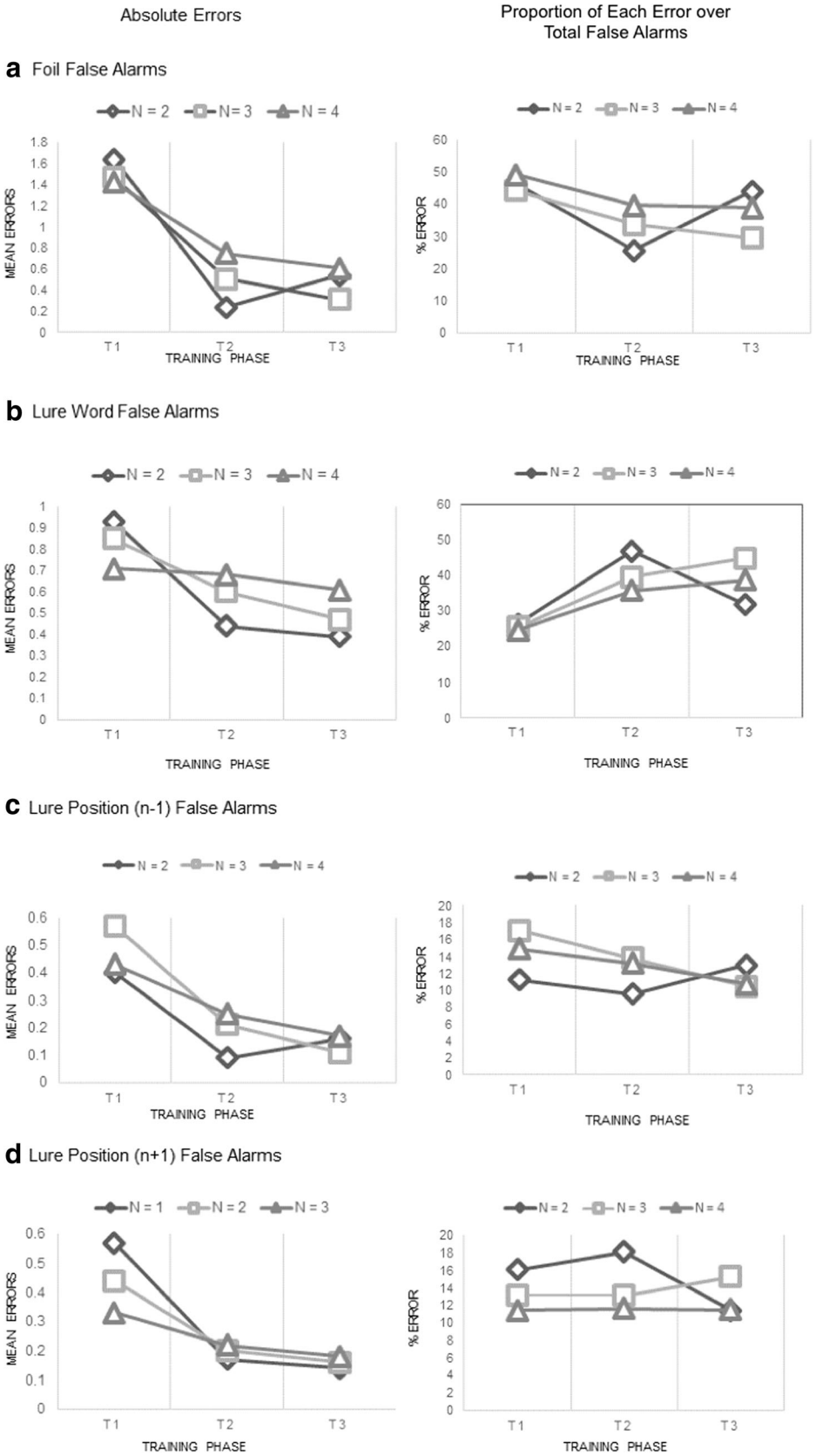

Fig. 9 Mean absolute errors and proportion of false alarms among Experiment 2 participants

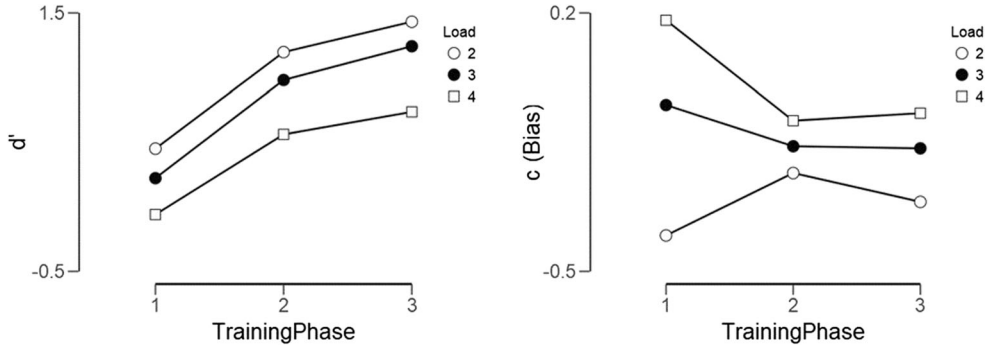

Fig. 10 Sensitivity index, *d'*, across three training phases (**left panel**) and Bias, *c*, across three training phases (**right panel**) among Experiment 2 participants

dependent variable. Baseline performance of reasoning ability and working-memory capacity were also analyzed as predictors for average load attained at sessions 1 and 15.

Results

Participants showed improvements on N-back after 15 sessions of training. There was a significant increase of mean load attempted from session 1 (mean = 2.98, SD = 0.63) to session 15 (mean = 5.96, SD = 1.16; t (19) = 11.00, p <0.001; BF_{10} = 9.73 x 10^6).

We summarized data analyses from loads 2, 3, and 4. Overall, total errors including targets missed and total false alarms decreased over time (Fig. 11). Results from ANOVA repeated measures and Bayesian repeated measures of the whole sample were summarized in Table 4. Mean errors of each type of false alarm reduced over time (Fig. 12, left column). The change in proportion of each type of false alarm (foil, lure word, lure position n-1, and lure position n+1) over total false alarms showed that foil false alarms decreased over time but false alarms due to lure word and lure position (n-1) did not decrease. False alarms due to lure position (n+1) showed a slight increase over time (Fig. 12, right column).

Descriptive statistics of all cognitive measures assessed in the study are presented in Table 5. Participants in the training group improved significantly on the symmetry span (t(19) = 3.28; p = 0.004; BF_{10} = 11.28). After removing three participants from the whole sample whose accuracy level fell below 80% (an indication of whether participants were or were not performing both processing and storage tasks equally well), there was still a significant difference in symmetry span performance in the training group (t(17) = 3.03; p = 0.008; BF_{10} = 6.61). There was no significant impact of training on any of the other measures, replicating results from Experiment 2.

A regression model with only reasoning ability measured by RAPM significantly predicted average load achieved at session 1 (B = 0.48, t = 2.50, p = 0.021; BF_{10} = 2.06) and session 15 (B = 0.47, t = 2.25, p = 0.038; BF_{10} = 3.05). Regression models that included working-memory capacity (OSPAN and symmetry span) do not significantly predict performance of N-back at the beginning or end of training.

Strategy reports

Self-report of strategies employed during training and frequency of different strategies used throughout training were analyzed using regression analyses and repeated-measures

Table 3 Pre- and post-training performance on each of the criterion task for participants in the n-back and OSPAN training group

Criterion Tasks	N-back training (n=29)		OSPAN training (n=26)	
	Pre	Post	Pre	Post
Letter Span – Distinct	6.00 (0.89)	6.38 (1.35)	6.12 (1.36)	6.54 (1.25)
Letter Span – Similar	6.14 (1.16)	6.31 (1.69)	6.35 (1.72)	6.32 (1.68)
Digit Span – Forward	7.72 (1.41)	8.28 (1.44)	7.27 (1.76)	7.75 (1.57)
Digit Span – Backward	7.50 (1.53)	7.55 (1.76)	6.92 (1.55)	7.08 (1.73)

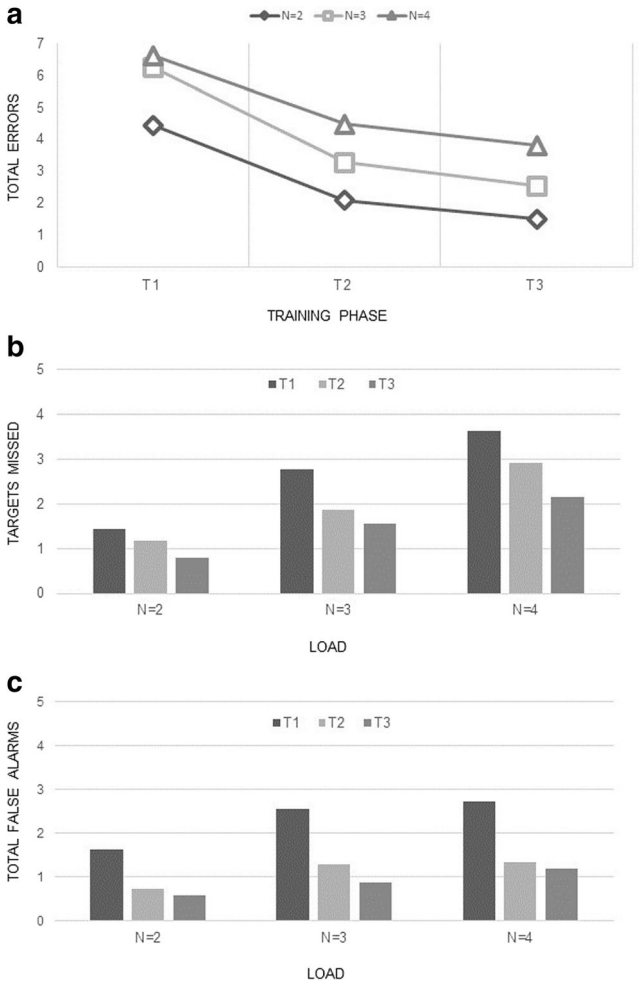

Fig. 11 Trends of total errors over three training phases (15 sessions) among participants in Experiment 3. (**A**) Total errors; (**B**) targets missed; (**C**) total false alarms

ANOVA. We predicted that employing complex strategies such as imagery and elaboration would lead to higher N (load) achieved at the end of training. None of the strategies reported (rehearsal, chunking, imagery, etc.; see Table 6) significantly predicted final average N (load) achieved at session 15 nor performance difference between sessions 1 and 15. Participants who reported changing strategies and employing a more complex and personalized strategy over the course of training attained a higher load (N) at the end of training compared to those who consistently maintained the same strategies throughout training, but the difference was not significant.

Repeated-measures ANOVA with strategies Rehearsal, Chunking, Chaining, Imagery and Elaboration as dependent variables indicated a significant main effect of training phase ($F(2) = 70.5$; $p < 0.001$; $BF_{10} = 290.7$), main effect of strategy ($F(4) = 277.4$, $p < 0.001$; $BF_{10} = 5.10 \times 10^{71}$), and interaction between strategy use and training phase ($F(8) = 51.3$, $p < 0.001$; $BF_{10} = 1.64 \times 10^{36}$; see Fig. 13A). Post hoc analyses indicated no significant change in Chaining and Imagery strategy use, so these strategies were removed from subsequent analyses. Strategies created by participants also showed a significant increase in use for "Slot Machine" and Acoustic strategies (see Fig. 13B). There was a significant main effect of training phase ($F(2) = 155.1$; $p < 0.001$; $BF_{10} = 5.26 \times 10^{11}$), main effect of strategy ($F(4) = 269.4$, $p < 0.001$; $BF_{10} = 1.22 \times 10^{57}$), and interaction between strategy use and training phase ($F(8) = 35.3$, $p < 0.001$; $BF_{10} = 4.88 \times 10^{23}$). Figure 13B suggested a significant interaction between strategy use and training phase for strategies Rehearsal, Chunking, Slot Machine, and Acoustic. These strategies were added as covariates in a repeated-measures ANOVA with lure word false alarms as dependent variable to evaluate if strategy use had an influence on rate of lure word false alarms. There were no

Table 4 Summary of repeated-measures ANOVA and Bayesian repeated-measures ANOVA of total errors and types of errors generated during n-back training sessions (Experiment 3) for loads 2, 3 and 4

	Training phase main effect		Load main effect		Interaction Training Phase x Load	
	F-value ($F(2)$)	Bayesian Factor (BF)	*F*-value ($F(2)$)	Bayesian Factor (BF)	*F*-value ($F(4)$)	Bayesian Factor (BF)
Total errors	68.78**	4.05×10^{16}	40.46**	81.3×10^{4}	0.89	0.10
Targets missed	12.64**	4775.63	50.00**	2.34×10^{13}	2.97*	0.25
Foil false alarms	27.81**	1.72×10^{17}	7.88**	44.95	8.13**	0.31
Lure word false alarms	4.56*	4.10	14.09**	501.94	1.85	0.42
Lure position n-1 false alarms	21.90**	6.32×10^{10}	14.17**	87.51	3.79*	0.73
Lure position n+1 false alarms	19.89**	2.60×10^{8}	2.63	0.23	0.81	0.11

**$p < 0.001$

*$p < 0.05$

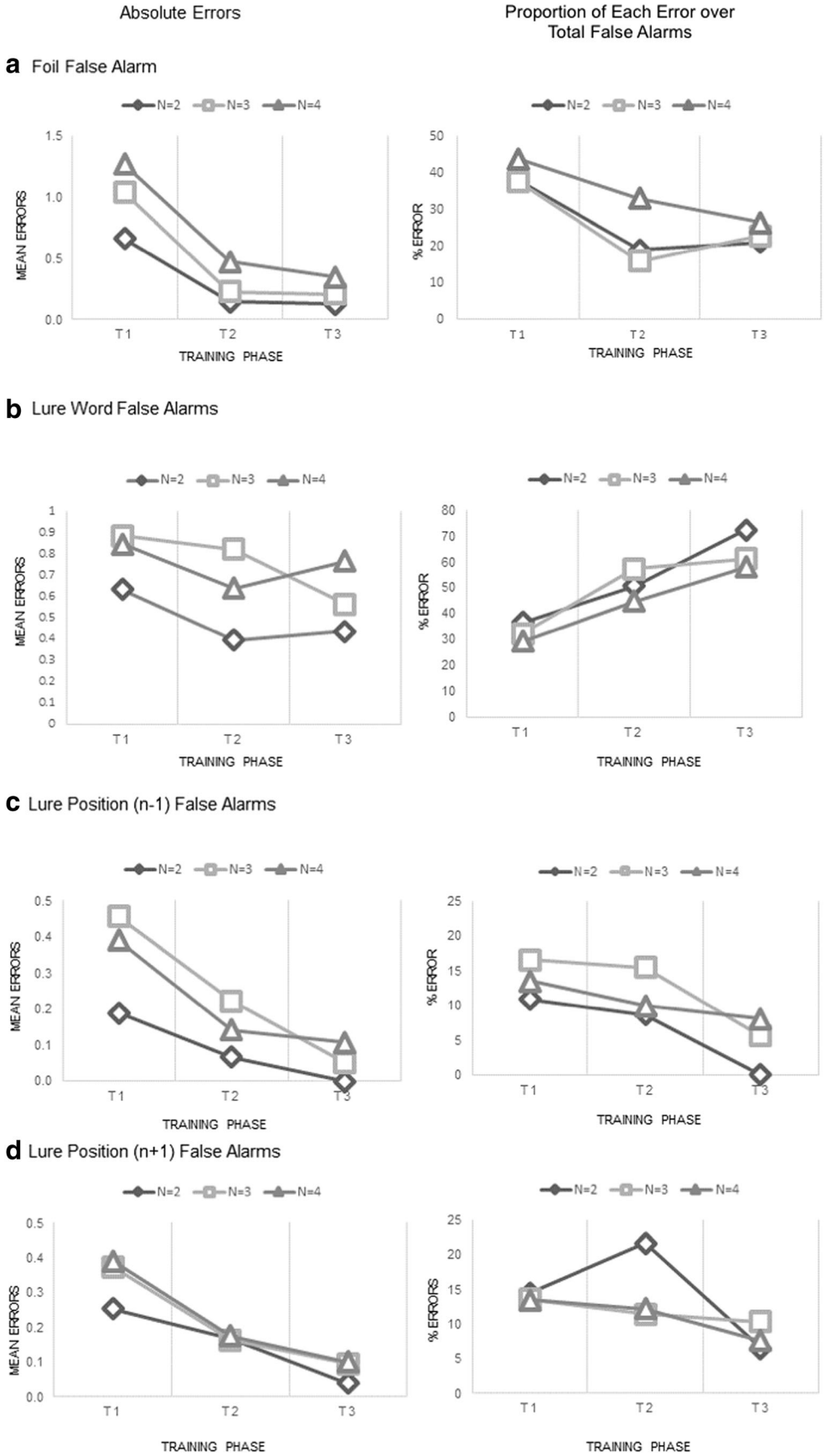

Fig. 12 Mean absolute errors and proportion of false alarms among Experiment 3 participants

significant within-subject and between-subject effects for strategy use in errors made due to lure words.

Item-level (word-pair) analysis

The eight pairs of homophone words selected were pilot-tested among non-native English speakers. Some of these word-pairs may not sound similar or be pronounced similarly to English native speakers. Repeated-measures one-way ANOVA of errors due to lure word false alarms were conducted (Table 7). In Experiments 1 and 3, participants made the least errors with the word-pairs "ate-eight," "feel-fill," and "pull-pool." In Experiment 2, participants made the least errors with the word-pairs "ate-eight" and "pull-pool." A between-group one-way ANOVA was conducted to test if there were differences in number of lure word false alarms across participants from all three experiments. Post hoc comparisons indicated significant group differences between Experiments 1 and 3 for word-pairs "blue-blew" ($F(2) = 3.58$, $p = 0.030$; $BF_{10} = 1.14$) and "soul-sole" ($F(2) = 4.32$, $p = 0.014$; $BF_{10} = 2.18$). Participants in Experiment 2 behaved significantly differently compared to participants in Experiments 1 and 3 for word-pair "feel-fill" ($F(2) = 4.60$, $p = 0.011$; $BF_{10} = 2.77$). They made significantly higher errors due to this word-pair compared to those in Experiments 1 and 3.

Discussion

In Experiment 3 with English native speakers in the United Kingdom, we replicated changes in error patterns during N-Back training observed in Experiments 1 and 2 conducted in Malaysia. Most importantly, the increase in proportion of errors due to homophone lure words was observed in all three experiments. All errors decreased over time but the rate of decrease for lure word errors was not as rapid as for the others. In all three samples, participants made fewer non-target-word (foil) false alarms that translated to improved performance overall as training progressed. However, proportion of false alarms due to phonologically similar lure words not only did not decrease but increased as training progressed (even though overall performance improved), suggesting that participants still heavily depended on sub-vocal rehearsal or auditory representation of stimuli. When asked how they performed the N-back, participants reported frequently engaging in sub-vocal rehearsal, which showed increased frequency over time ($F(2) = 9.55$, $p < 0.001$; $BF_{10} = 168.5$). A few participants reported trying to reduce phonologically similarity between homophone words by sub-vocally pronouncing some of the words in different ways, which provided further support that participants relied on phonological processing when training on the N-back task. In a task that demands heavy attention to monitoring and comparing the current stimulus with the n prior stimulus, participants adopt a simpler maintenance mechanism such as the articulatory rehearsal instead of attentional refreshing. Attentional refreshing is preferred or adopted in tasks that are less attentional demanding because attentional resources can be directed to non-phonological characteristic of memory items such as their semantic and visual features (Camos, Mora, & Oberauer, 2011). However, frequency with which participants reported a specific strategy did not significantly predict improved performance on the N-back after 3 weeks (15 sessions) of training. The lack of a clear relationship between the observed error patterns and reported

Table 5 Baseline and post-training performance of all cognitive tasks administered

Cognitive tests	Training group (N=20)		Control group (N=12)	
	Pre	Post	Pre	Post
Raven's Advanced Progressive Matrices	11.45 (2.53)	11.80 (2.12)	11.17 (2.76)	10.6 (2.54)
Operation Span	20.1 (4.12)	21.4 (4.08)	17.8 (6.78)	18.9 (6.99)
Symmetry Span	***18.4*** (6.30)	***21.2*** (4.51)	20.8 (4.81)	20.1 (5.96)
Letter Span Distinct	7.20 (1.32)	7.60 (1.88)	6.58 (1.17)	6.58 (1.62)
Letter Span Similar	6.70 (1.72)	7.35 (1.90)	6.08 (1.24)	6.25 (2.01)
Digit Span Forward	8.15 (1.27)	8.10 (1.25)	7.25 (2.01)	7.50 (1.00)
Digit Span Backward	7.25 (1.80)	7.60 (1.47)	7.00 (1.86)	7.00 (1.76)

Bolded and italicized values indicate significant paired-sample difference between pre- and post-training performance at p value < 0.001

Table 6 List of strategies used by participants during training. Italicized items were provided by the researcher at session 5 (time 1). Non-italicized items were provided by participants themselves

Strategy category	Item
Rehearsal	*I repeat out loud the words as they appear*
	I repeat mentally (sub-vocally) the words as they appear
	I repeat and rehearse the words as they appear
	I repeat each sequence of letters / words before making a response
Chunking	*I group the words in strings of 2 or 3 words*
	I group the words in strings of n words according to the level of n-back I am doing
	When doing 6-back and higher, I mentally group only the last 4 words
	When doing 7-back and higher, I mentally group words into sets of 4
Chaining	*I link or connect the words to form a story*
Imagery	*I imagine / create a mental picture of each word*
	I try to form a visual image / story of the words that appear
	I group words in twos and threes and make stories / visual images with these groupings whenever possible
	I picture the words in rows to cross-check them (compare whether the words in the same column match)
Elaboration	*I associate each word with something I am familiar with in my knowledge database*
	I remember the first few letters of the words and form them into words or acronyms I am familiar with
"Slot Machine"	I remember the number of words according to N-back and mentally replace the Nth word with every new word
	I create N number of slots in my mind according to N-back and fill up the slots with words that appear – consciously move words backward
Acoustic	I differentiate homophones by adding suffixes to words (e.g., here vs. hearing)
	I remember the first few letters of the words when remembering them (e.g., stress the "oo" in "pool"
	I group the words into tones to form a song in my head and imagine they are song lyrics / poetry
	For 6- and 7-back, I tried using tunes to remember the sequence of words
	I mentally mispronounce some words like "sale" to distinguish it from "soul"
Kinaesthetic	I tap / count with fingers to track how many words to remember
	I do actions with homophones (e.g., blew - blow air; move my arm - vein; tap my foot - sole)

strategies suggests that the subjective reports of strategies should be treated with caution.

Changes in criterion tasks after training

There was no significant difference between pre- and post-training performance in any of the criterion tasks assessed except for symmetry span. In a recent publication proposing the cognitive routine framework in adaptive training studies, transfer was only possible if new cognitive routines had to be established when performing a highly unfamiliar task (Gathercole et al., 2019). The authors proposed that when one repeatedly practiced on a task, it is likened to acquiring a new set of complex cognitive routines that coordinates

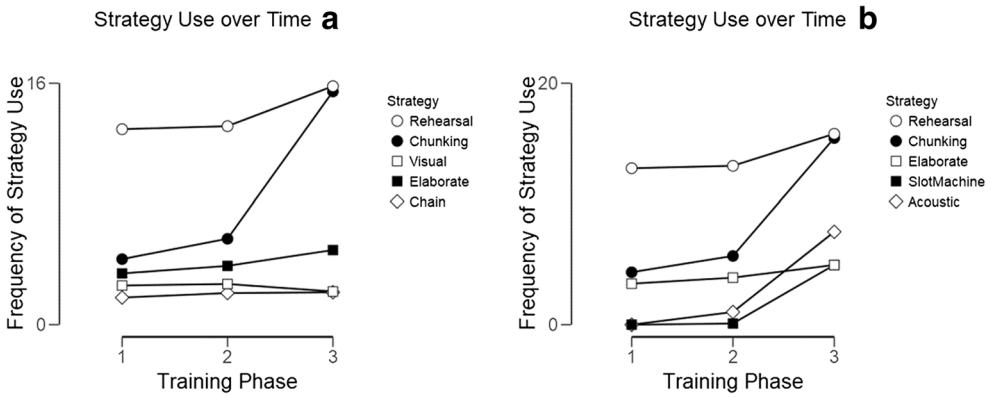

Fig. 13 Frequency of various strategy use over the course of N-back training in Experiment 3

Table 7 Descriptive statistics for each word-pair in all three experiments

	Experiment 1 Mean (SD)	Experiment 2 Mean (SD)	Experiment 3 Mean (SD)
Ate-eight	2.35 (2.05)	3.35 (2.76)	3.11 (1.78)
Blue-blew	4.18 (2.80)	4.47 (4.34)	5.66 (3.25)
Feel-fill	3.77 (3.10)	4.91 (3.73)	3.48 (2.21)
Hear-here	5.66 (3.43)	5.82 (3.40)	6.52 (3.57)
Made-maid	4.88 (3.37)	5.75 (4.36)	5.72 (3.23)
Pool-pull	3.80 (2.72)	4.60 (3.17)	4.17 (2.45)
Sole-soul	4.88 (3.27)	5.59 (3.97)	6.91 (4.50)
Vain-vein	5.18 (3.47)	5.97 (4.50)	6.66 (3.87)

multiple existing cognitive subroutines. The cognitive routine framework explained and provided support that if the learned or acquired new cognitive routine can be applied to untrained tasks after training, transfer would occur (Gathercole et al., 2019). The authors also explained that transfer was not as simple as matching task elements (e.g., same domain stimulus, response modality, etc.) between the trained and untrained tasks. Their meta-analysis suggested that when the untrained and trained tasks employed well-established cognitive routines such as rehearsal and recall of serial order of verbal stimuli, transfer effects were minimal because training only fine-tuned those routines and therefore training of such routines did not produce substantial transfer effects.

Adopting the cognitive routine framework in interpreting our results, we suggest that the performance on symmetry span improved after adaptive N-back training because the complex cognitive routine acquired during training was adopted in executing the symmetry span. In symmetry span, participants alternated between judging symmetrical figures and remembering spatial locations. Gathercole et al. (2019) suggested that training paradigms that lead to developing a routine to remember visual-spatial stimuli showed more robust transfer effects to untrained tasks that would benefit from applying the same cognitive routine in performing the tasks because the visual-spatial short term memory (STM) was not as well established compared to verbal STM in most people. We suggest that participants in our study acquired a new complex cognitive routine while training on the adaptive N-back and this cognitive routine helped them in remembering serial order spatial stimuli even when interfered by another visual processing task. However, precisely what this cognitive routine might be, and why it would arise from visual N-back training is unclear. The impact that we observed on symmetry span was small in absolute score, even if significant, and so it would be important to investigate this possible effect of N-back training on symmetry span to ensure that it replicates in future studies

The cognitive routine framework also proposed that establishing a new complex cognitive routine would draw on general cognitive resources outside of working memory. A few studies have observed that participants with high general cognitive ability at baseline gained more improvements in training studies (Foster et al., 2017; S. M. Jaeggi et al., 2008), which led to the speculation that individuals with more abilities to generate multiple strategies were able to improve more rather than the strategy itself causing improvements. In our study, reasoning ability as measured at baseline by RAPM significantly predicted N-back performance at the first and last training sessions. Working memory capacity as measured by complex span tasks, specifically OSPAN, did not significantly predict N-back performance.

General discussion

Our first experiment indicated that, over time, participants were able to detect more targets by relying more on phonological coding in remembering or holding items in their immediate memory. As training progressed, this was reflected by an increase in proportion of false alarms due to lure words that were phonologically similar to target words. These results were replicated in Experiment 2, although none of our predictions of improved recall in simple span tasks after training were supported.

We were able to replicate error patterns from Experiments 1 and 2, conducted in Malaysia, in Experiment 3, conducted in the United Kingdom. Data from two international regions confirmed that participants became more reliant on phonological processing of task stimuli as they trained on a visual N-back task. In the analysis of self-report strategy use across the whole group during training, we observed that reported strategy use did not influence N-back performance. However, reasoning ability at baseline as measured by RAPM significantly predicted N-back performance, whereas working-memory capacity as measured by OSPAN did not.

Summary and conclusions

The aim of our study was to enhance the theoretical understanding of changes in participants' cognition during cognitive training using an adaptive N-back task, and whether that understanding can help predict and provide an explanation for any outcomes of the training that transferred or did not transfer to performance on other tasks. Across three experiments we discovered that participants relied more heavily on phonological coding of items that were held in working memory as they trained, and this was associated with improved performance on the task on which they trained. No transfer effects were detected in our study.

In summary, we have reported a novel finding based on error patterns that during N-back training participants increasingly use phonologically based rehearsal and change their response criteria. However, changes in how the task is performed as training progresses does not reliably transfer to performance on tasks that are also thought to rely on phonologically based rehearsal. These results reinforce conclusions from some previous studies suggesting that cognitive training results in improvements on the trained task itself with only limited or no reliable evidence for transfer to other tasks (e.g., Chooi & Thompson, 2012; De Simoni & von Bastian, 2018; Redick et al., 2013; Shipstead et al., 2012; Simons et al., 2016). Our study adds novel insight into possible changes in cognition during training when participants select their own cognitive strategies to support task performance. Contrasting findings from previous studies might be resolved by exploring in more depth whether training benefits are observed or not because different strategies are adopted during training by participants across different laboratories (for a general discussion see Logie, 2018). Future studies might adopt the approach of instructing participants to use specific strategies during training (e.g., Laine et al., 2018), rather than leaving strategy development to vary according to participant preferences. Selection of the instructed strategies might also be based on the Gathercole et al. (2019) proposal that training benefits accrue from learning new cognitive subroutines rather than practicing established strategies such as sub-vocal rehearsal. This approach would capitalize on what is already well known about the effects of practice on task performance, and offer a more systematic approach to predicting what other cognitive tasks might or might not benefit from that practice.

Acknowledgements This research was supported by the Newton-Ungku Omar Fund administered by the British Academy and Akademi Sains Malaysia under Newton Advanced Fellowship (AF 160093).

Open practices statement The data and materials for all experiments are available upon request, and none of the experiments were preregistered.

References

Au, J., Buschkuehl, M., Duncan, G. J., & Jaeggi, S. M. (2016). There is no convincing evidence that working memory training is NOT effective: A reply to Melby-Lervåg and Hulme (2015). *Psychonomic Bulletin & Review*, *23*(1), 331–337. https://doi.org/10.3758/s13423-015-0967-4

Bailey, H. R., Dunlosky, J., & Hertzog, C. (2014). Does Strategy Training Reduce Age-Related Deficits in Working Memory? *Gerontology*, *60*(4), 346–356. https://doi.org/10.1159/000356699

Bottiroli, S., Cavallini, E., & Vecchi, T. (2008). Long-term effects of memory training in the elderly: A longitudinal study. *Archives of Gerontology and Geriatrics*, *47*(2), 277–289. https://doi.org/10.1016/J.ARCHGER.2007.08.010

Bunting, M., Cowan, N., & Saults, J. S. (2006). How does running memory span work? *Quarterly Journal of Experimental Psychology (2006)*, *59*(10), 1691–1700. https://doi.org/10.1080/17470210600848402

Camos, V., Mora, G., & Oberauer, K. (2011). Adaptive choice between articulatory rehearsal and attentional refreshing in verbal working memory. *Memory & Cognition*, *39*(2), 231–244. https://doi.org/10.3758/s13421-010-0011-x

Chooi, W.-T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*(6), 531–542. https://doi.org/10.1016/j.intell.2012.07.004

Colley, A. M., & Beech, J. R. (Eds.). (1989). *Acquisition and Performance of Cognitive Skills*. Chichester: John Wiley & Sons.

Contreras, N. A., Tan, E. J., Lee, S. J., Castle, D. J., & Rossell, S. L. (2018). Using visual processing training to enhance standard cognitive remediation outcomes in schizophrenia: A pilot study. *Psychiatry Research*, *262*, 494–499. https://doi.org/10.1016/j.psychres.2017.09.031

Dardiotis, E., Nousia, A., Siokas, V., Tsouris, Z., Andravizou, A., Mentis, A.-F. A., … Nasios, G. (2018). Efficacy of computer-based cognitive training in neuropsychological performance of patients with multiple sclerosis: A systematic review and meta-analysis. *Multiple Sclerosis and Related Disorders*, *20*, 58–66. https://doi.org/10.1016/j.msard.2017.12.017

De Luca, R., Leonardi, S., Spadaro, L., Russo, M., Aragona, B., Torrisi, M., … Calabrò, R. S. (2018). Improving Cognitive Function in Patients with Stroke: Can Computerized Training Be the Future? *Journal of Stroke and Cerebrovascular Diseases*, *27*(4), 1055–1060. https://doi.org/10.1016/j.jstrokecerebrovasdis.2017.11.008

De Simoni, C., & von Bastian, C. C. (2018). Working memory updating and binding training: Bayesian evidence supporting the absence of transfer. *Journal of Experimental Psychology: General*, *147*(6), 829–858. https://doi.org/10.1037/xge0000453

Donchin, E., Fabiani, M., & Sanders, A. (1989). The Learning Strategies Program: An examination of the strategies in skill acquisition. *Acta Psychologica*, *71*, 1–312.

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

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working-memory capacity. *Memory & Cognition* https://doi.org/10.3758/s13421-014-0461-7

Foster, J. L., Harrison, T. L., Hicks, K. L., Draheim, C., Redick, T. S., & Engle, R. W. (2017). Do the effects of working memory training depend on baseline ability level? *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Engle, Randall W.: School of Psychology, Georgia Institute of Technology, 654 Cherry Street, Atlanta, GA, US, randall.engle@gatech.edu: American Psychological Association. 10.1037/xlm0000426

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

Genevsky, A., Garrett, C. T., Alexander, P. P., & Vinogradov, S. (2010). Cognitive training in schizophrenia: a neuroscience-based approach. *Dialogues in Clinical Neuroscience*, *12*(3), 416–421. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3181983/

Han, K., Chapman, S. B., & Krawczyk, D. C. (2018). Neuroplasticity of cognitive control networks following cognitive training for chronic traumatic brain injury. *NeuroImage: Clinical*, *18*, 262–278. https://doi.org/10.1016/j.nicl.2018.01.030

Jaeggi, S., Buschkuehl, M., Perrig, W., & Meier, B. (2010). The concurrent validity of the N -back task as a working memory measure. *Memory (Hove, England)*, *18*, 394–412. https://doi.org/10.1080/09658211003702171

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(19), 6829–6833. https://doi.org/10.1073/pnas.0801268105

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The Generality of Working Memory Capacity: A Latent-Variable Approach to Verbal and Visuospatial Memory Span and Reasoning. *Journal of Experimental Psychology: General*. Kane, Michael J.: Department of Psychology, University of North Carolina at Greensboro, P.O. Box 26170, Greensboro, NC, US, 27402-6170, mjkane@uncg.edu: American Psychological Association. https://doi.org/10.1037/0096-3445.133.2.189

Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal of Clinical and Experimental Neuropsychology*, *24*(6), 781–791. https://doi.org/10.1076/jcen.24.6.781.8395

Laine, M., Fellman, D., Waris, O., & Nyman, T. J. (2018). The early effects of external and internal strategies on working memory updating training. *Scientific Reports*, *8*(1), 4045. https://doi.org/10.1038/s41598-018-22396-5

LIU, X. Y., LI, L., XIAO, J. Q., HE, C. Z., LYU, X. L., GAO, L., … FAN, L. H. (2016). Cognitive Training in Older Adults with Mild Cognitive Impairment. *Biomedical and Environmental Sciences*, *29*(5), 356–364. https://doi.org/10.3967/bes2016.046

Logie, R. (2018). Human cognition: Common Principles and Individual Variation. *Journal of Applied Research in Memory and Cognition*, *7*, 471–486. https://doi.org/10.1016/j.jarmac.2018.08.001

Logie, R. H. (2012). Cognitive training: Strategies and the multicomponent cognitive system. *Journal of Applied Research in Memory and Cognition*, *1*(3), 206–207. https://doi.org/10.1016/j.jarmac.2012.07.006

Logie, R. H., Baddeley, A. D., Mane, A., Donchin, E., & Sheptak, R. (1989). Working memory and the analysis of a complex skill by secondary task methodology. *Acta Psychologica*, *71*, 53–87.

McBride, R. L., Horsfield, S., Sandler, C. X., Cassar, J., Casson, S., Cvejic, E., … Lloyd, A. R. (2017). Cognitive remediation training improves performance in patients with chronic fatigue syndrome. *Psychiatry Research*, *257*, 400–405. https://doi.org/10.1016/j.psychres.2017.08.035

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

Melby-Lervåg, M., Redick, T., & Hulme, C. (2016). Working Memory Training Does Not Improve Performance on Measures of Intelligence or Other Measures of "Far Transfer": Evidence from a Meta-Analytic Review. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *11*(4), 512–534. https://doi.org/10.1177/1745691616635612

Peng, P., & Fuchs, D. (2015). A Randomized Control Trial of Working Memory Training With and Without Strategy Instruction: Effects on Young Children's Working Memory and Comprehension. *Journal of Learning Disabilities*, *50*(1), 62–80. https://doi.org/10.1177/0022219415594609

Raven, J. C., Court, J. H., & Raven, J. (1977). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. London: H.K. Lewis & Co. Ltd.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., … Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*. Redick, Thomas S.: Division of Science, Indiana University-Purdue University Columbus, 4601 Central Avenue, Columbus, IN, US, 47203, tsredick@iupuc.edu: American Psychological Association. https://doi.org/10.1037/a0029082

Shipstead, Z., Hicks, K. L., & Engle, R. W. (2012). Cogmed working memory training: Does the evidence support the claims? *Journal of Applied Research in Memory and Cognition*, *1*(3), 185–193. https://doi.org/10.1016/j.jarmac.2012.06.003

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. L. (2016). Do "Brain-Training" Programs Work? *Psychological Science in the Public Interest*, *17*(3), 103–186. https://doi.org/10.1177/1529100616661983

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic Bulletin & Review*, *24*(4), 1077–1096. https://doi.org/10.3758/s13423-016-1217-0

Subramaniam, K., Gill, J., Fisher, M., Mukherjee, P., Nagarajan, S., & Vinogradov, S. (2018). White matter microstructure predicts cognitive training-induced improvements in attention and executive functioning in schizophrenia. *Schizophrenia Research*, *193*, 276–283. https://doi.org/10.1016/j.schres.2017.06.062

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, *8*, 247–261.

Towe, S. L., Patel, P., & Meade, C. S. (2017). The Acceptability and Potential Utility of Cognitive Training to Improve Working Memory in Persons Living With HIV: A Preliminary Randomized Trial. *Journal of the Association of Nurses in AIDS Care*, *28*(4), 633–643. https://doi.org/10.1016/j.jana.2017.03.007

Unsworth, N., & Engle, R. W. (2006). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language*, *54*(1), 68–80. https://doi.org/10.1016/j.jml.2005.06.003

Vogt, A., Kappos, L., Calabrese, Pasquale Stöcklin, M., Gschwind, L., Opwis, K., & Penner, I.-K. (2009). Working memory training in patients with multiple sclerosis – comparison of two different training schedules. *Restorative Neurology and Neuroscience*, *27*(3), 225–235. https://doi.org/10.3233/RNN-2009-0473

Westerberg, H., Jacobaeus, H., Hirvikoski, T., Clevberger, P., Östensson, M.-L., Bartfai, A., & Klingberg, T. (2007). Computerized working memory training after stroke–A pilot study. *Brain Injury*, *21*(1), 21–29. https://doi.org/10.1080/02699050601148726

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.