Intelligence xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Intelligence

ELSEVIER

Improvement in working memory is not related to increased intelligence scores

Roberto Colom [a,*], Mª Ángeles Quiroga [b], Pei Chun Shih [a], Kenia Martínez [a], Miguel Burgaleta [a], Agustín Martínez-Molina [a], Francisco J. Román [b], Laura Requena [a], Isabel Ramírez [b]

[a] Universidad Autónoma de Madrid, Spain
[b] Universidad Complutense de Madrid, Spain

ARTICLE INFO

Article history:
Received 28 December 2009
Received in revised form 12 June 2010
Accepted 15 June 2010
Available online xxxx

Keywords:
Cognitive improvements
Intelligence
Short-term memory
Working memory
Processing speed
Attention

ABSTRACT

The acknowledged high relationship between working memory and intelligence suggests common underlying cognitive mechanisms and, perhaps, shared biological substrates. If this is the case, improvement in working memory by repeated exposure to challenging span tasks might be reflected in increased intelligence scores. Here we report a study in which 288 university undergraduates completed the odd numbered items of four intelligence tests on time 1 and the even numbered items of the same tests one month later (time 2). In between, 173 participants completed three sessions, separated by exactly one week, comprising verbal, numerical, and spatial short-term memory (STM) and working memory (WMC) tasks imposing high processing demands (STM–WMC group). 115 participants also completed three sessions, separated by exactly one week, but comprising verbal, numerical, and spatial simple speed tasks (processing speed, PS, and attention, ATT) with very low processing demands (PS-ATT group). The main finding reveals increased scores from the pre-test to the post-test intelligence session (more than half a standard deviation on average). However, there was no differential improvement on intelligence between the STM-WMC and PS-ATT groups.

1. Introduction

The consensus APA report on intelligence, chaired by Ulric Neisser and signed by eleven scientists and practitioners chosen to represent different psychological approaches, accepts that improvements in intelligence are possible, but it also acknowledges that observed "*gains fade when the program is over*" (Neisser, Boodoo, Bouchard, Boykin, Brody, Ceci et al., 1996, p. 96). The mainstream science on intelligence report, chaired by Linda Gottfredson and signed by 52 scientists, summarizes the available evidence regarding attempts to raise intelligence: "*although the environment is important in creating intelligence differences, we do not know yet how to manipulate it to raise low intelligence permanently*" (Gottfredson et al., 1997, p. 15). The treatise published by A. R. Jensen (*The g factor*) reviews programs aimed at raising intelligence. In his own words "*few other topics in the history of behavioral science have resulted in so vast literature, and no other comes close to it in total research expenditure*" (Jensen, 1998, p. 333). The main conclusion of Jensen's review is that we know how to improve certain skills, but we don't know how to improve general intelligence (*g*).

One of the main reservations regarding the numerous attempts made for improving intelligence is related to the coaching phenomenon (Brody, 1992; Brown, & Campione, 1982). Participants on an improvement program may show increased performance on particular mental tests, but transfer or generalizability usually fails. From this perspective it is especially important to make sure that training methods and ways for assessing improvements in intelligence are clearly distinguishable.

* Corresponding author. Facultad de Psicología, Universidad Autónoma de Madrid, 28049 Madrid, Spain. Tel.: +34 91 497 41 14.
E-mail address: roberto.colom@uam.es (R. Colom).

0160-2896/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2010.06.008

This is the general approach endorsed in the report by Jaeggi, Buschkuehl, Jonides, and Perrig (2008). These researchers have shown that improvements in working memory performance are related to higher fluid intelligence scores. This study identified a cognitive task thought to share relevant underlying mental processes with fluid intelligence tests. Importantly, the cognitive task was different enough to avoid simple coaching.

The main framework for this study is based on Halford, Cowan, and Andrews (2007) hypothesis: working memory and intelligence share common capacity limitations. The limitations come from the number of items that can be kept active in working memory or the number of relationships between elements that can be kept active during the reasoning process necessary for solving problems comprised on standard intelligence tests. The latter impose processing loads also required by working memory tasks. These shared limitations could be based on the ability to build and keep bindings among items in the short-term. A similar view has been proposed by Kyllonen, and Christal (1990) or by Colom, Abad, Quiroga, Shih, and Flores-Mendoza (2008), Colom, Abad, Rebollo, and Shih (2005), Colom, Rebollo, Palacios, Juan-Espinosa, and Kyllonen (2004) using correlational approaches, as well as on the Carpenter, Just, and Shell (1990) seminal study of the Advanced Progressive Matrices Test.

Jaeggi et al. (2008) developed a training paradigm based on the working memory *n* back task. In the modeled task, participants see two sets of stimuli synchronously presented at the rate of 3 s per stimulus. The first set consists of letters and the second set consists of spatial locations. Participants must decide for each set whether the current stimulus match the one presented *n* items back. The value for *n* changes from block to block depending on participants' performance. Therefore, this dual *n* back task is always demanding for each participant and high level control processes are systematically involved. Participants were tested to obtain their intelligence scores before and after training on the dual *n* back task. In order to control for retest effects, performance of the trained group was compared with a non-trained group.

The main findings showed higher intelligence posttest scores for the trained group. This was interpreted after the nature of the *n* back task modeled in the study, which engages several diverse executive processes: inhibition of irrelevant items, monitoring ongoing performance, managing two tasks at the same time, updating representations in working memory, as well as binding processes between letters and spatial locations and their temporal framework. However, giving this interpretation it is difficult to explain why the trained and control groups did not differ on a posttest administration of a standard measure of working memory (reading span). The authors were inclined to endorse an interpretation based on improvements in the control of attention: "*one reason for having obtained transfer between working memory and measures of fluid intelligence is that our training procedure may have facilitated the ability to control attention*" (p. 6831).

Moody (2009) critically reviewed this study noting that Jaeggi et al. had indeed four trained groups. One group was tested using the Raven test, whereas the remaining three groups were tested using the BOMAT test. Observed improvements were circumscribed to the second test. However, the BOMAT test was administered with a time limit of 10 min instead of the standard 45 min. Therefore, it was impossible for the subjects to face the more difficult items, so, perhaps, the test no longer challenged their fluid intelligence. What this rare administration of the BOMAT test challenges is the speed by which participants can manipulate simultaneously the 5 × 3 matrix configurations of simple perceptual items. Moody concludes: "*whatever the meaning of the modest gains in performance on the BOMAT, the evidence produced by Jaeggi et al. does not support the conclusion of an increase in their subjects' intelligence. Their research may be sufficient to encourage further investigation, but any larger inferences are unwarranted*" (p. 328).

The present study is intended to contribute to this further investigation. Here we analyze performance changes of two hundred and eighty eight participants who completed several testing sessions across five weeks. In the first week, the whole sample was tested using the odd items of four intelligence tests: the Advanced Progressive Matrices Test (APM), along with the Abstract Reasoning, Verbal Reasoning, and Spatial Relations subtests from the Differential Aptitude Test Battery. Afterwards, the sample was divided in two groups. 173 participants performed three short-term (forward letter span, forward digit span, and corsi block) and three working memory tasks (reading span, computation span, and dot matrix), whereas 115 participants performed three processing speed (verbal, numerical, and spatial recognition speed) and three attention tasks (verbal flanker, numerical flanker, and Simon tasks). These computerized short-term memory (STM), working memory (WMC), processing speed (PS), and attention (ATT) tasks were administered on weeks two, three, and four. Finally, on the fifth week, the whole sample completed the even items of the same intelligence tests administered on the first week.

Of note is that (a) STM and WMC tasks were administered for always engaging complex and effortful processing and (b) tap several diverse high-level cognitive processes because of their verbal, numerical, and spatial nature. Further, difficulty levels were randomly presented for every task and the order of administration was counter-balanced across the three weeks. PS and ATT tasks were very low cognitively demanding, but order of administration was also counter-balanced.

We predict generalized improvements in intelligence, STM, WMC, PS, and ATT. However, for reasons stated above, those participants that face cognitive tasks engaging high demanding working memory processes must show better post-test scores on the intelligence measures. It must be underscored that intelligence tests were administered using standard time limits.

2. Method

2.1. Participants

288 psychology undergraduates took part in the study (82% were females). The mean age was 20.1 (SD = 3.4). They participated to fulfill a course requirement. Participants were randomly assigned to two groups. The first group (STM-WMC group) was comprised by 173 participants, whereas the second group (PS-ATT group) was comprised by 115 participants.

It is hard to tell whether having such a great number of females is relevant for the present study. We are inclined to think that it is not, because, as it is shown in the next section, tests and tasks were largely heterogeneous with no remarkable verbal or spatial bias. Anyways, as requested by one reviewer, data for males and females are reported in the Appendix A.

2.2. Measures

Intelligence was measured by four tests: the Advanced Progressive Matrices Test (APM) along with the abstract reasoning (DAT-AR), verbal reasoning (DAT-VR), and spatial relations (DAT-SR) subtests from the Differential Aptitude Test Battery (Bennett, Seashore, & Wesman, 1990).

The APM (Set II) comprises a matrix figure with three rows and three columns with the lower right hand entry missing. There are eight alternatives and participants must choose the one completing the 3 × 3 matrix figure. Three items from Set I were administered as examples. The score was the total number of correct responses and total administration time was 40 min (20 min for the 18 odd items and 20 min for the 18 even items).

DAT-AR is a series test based on abstract figures. 40 items are comprised in this test. Each item includes four figures following a given rule, and the participant must choose one of five possible alternatives. The score was the total number of correct responses and total administration time was 20 min (10 min for the 20 odd items + 10 min for the 20 even items).

DAT-VR is a reasoning test comprising 40 items. A given sentence stated like an analogy must be completed. The first and last words from the sentence are missing, so a pair of words must be selected to complete the sentence from five possible alternative pairs of words. For instance: *….. is to water like eating is to ….. (A) Travelling–Driving, (B) Foot–Enemy, (C) Drinking–Bread, (D) Girl–Industry, (E) Drinking–Enemy*. Only one alternative is correct. The score was the total number of correct responses and total administration time was 20 min (10 min for the 20 odd items + 10 min for the even items).

Finally, DAT-SR is a mental folding test comprising 50 items. Each item is composed by an unfolded figure and four folded alternatives. The unfolded figure is shown at the left, whereas figures at the right depict folded versions. Participants are asked to choose one folded figure matching the unfolded figure at the left. The score was the total number of correct responses (well chosen folded figures) and total administration time was 20 min (10 min for the 25 odd items + 10 min for the 25 even items).

Short-term memory (STM) was measured by the forward letter span (FLSPAN), forward digit span (FDSPAN), and corsi block tasks. In the FLSPAN and FDSPAN tasks, single letters or digits (from 1 to 9) were presented on the computer screen at the rate of one letter or digit per 650 ms. Unlimited time was allowed to type in direct order the letters or digits presented. Set size of the experimental trials ranged from four to nine items (6 levels × 3 trials each = 18 trials total). Letters or digits were randomly grouped to form trials. Difficulty levels were randomly presented. The score was the number of accurately reproduced items (max = 117).

In the corsi block task nine boxes were shown on the computer screen. Three different configurations of boxes changing randomly on each trial were used to discourage perceptual strategies. One box at a time turned orange for 650 ms each and the order in which they were sequentially highlighted must be remembered. There was unlimited time to respond. The sequences of the experimental trials increased from 4 to 9 (6 levels × 3 trials each = 18 trials total). Difficulty levels were randomly presented. The score was the number of highlighted boxes reproduced appropriately (max = 117).

Working memory (WMC) was measured by the reading span, computation span, and dot matrix tasks. These WMC tasks were modified after Colom, Abad, Quiroga, Shih, and Flores-Mendoza (2008). In the reading span task participants verified which discrete sentences, presented in a sequence, did or did not make sense. Sentences were adapted from the Spanish standardization of the Daneman, and Carpenter (1980) reading span test (Elosúa, Gutiérrez, García-Madruga, Luque, & Gárate, 1996). Each display included a sentence and a to-be remembered capital letter. Sentences were 10–15 words long. As soon as the sentence–letter pair appeared, participants verified whether it did or did not make sense (it did half the time) reading the capital letter for latter recall. Once the sentence was verified by pressing the answer buttons (yes/1–no/2) the next sentence–letter pair was presented. At the end of a given set, participants recalled, in their correct order in the alphabet, each letter from the set. Set sizes of the experimental trials ranged from 3 to 7 sentence/letter pairs per trial, for a total of 15 trials (5 levels × 3 trial = 15 trials total). Difficulty levels were randomly presented. The score was the number of correct answers in the verification and recalling tasks (max = 150).

The computation span task included a verification task and a recall task. 6 s were allowed to see a math equation (but no time limit was set to verify its accuracy) like (10/2) + 4 = 8, and the displayed solution, irrespective of its accuracy, must be remembered. After the final equation of the trial was displayed, the solutions from the equations must be reproduced in their correct serial order. Each math equation included two operations using digits from 1 to 10. The solutions were single-digit numbers. The experimental trials ranged from three to seven equation/solutions (5 levels × 3 trials each = 15 trials total). Difficulty levels were randomly presented. The score was the number of correct answers in the verification and recalling tasks (max = 150).

In the dot matrix task a matrix equation must be verified and then a dot location displayed in a five × five grid must be retained. The matrix equation required adding or subtracting simple line drawings and it was presented for a maximum of 4.5 s. Once the response was delivered, the computer displayed the grid for 1.5 s. After a given sequence of equation-grid pairs, the grid spaces that contained dots must be recalled clicking with the mouse on an empty grid. The experimental trials increased in size from three to five equations and dots (3 levels × 3 trials = 9 trials total). Difficulty levels were randomly presented. The score was the number of correct answers in the verification and recalling tasks (max = 72).

Verbal, quantitative, and spatial processing speed (PS) was measured by simple verification tasks (short-term recognition speed). Participants were requested to verify, as quickly and accurately as possible, if a given test stimulus was presented previously. The participants pressed the computer key 1 for a "yes" answer and the computer key 0 for a "no" answer.

In the verbal speed task, one letter was displayed for 650 ms. After this presentation, a fixation point appeared for

500 ms. Finally, the probe letter appeared in order to decide, as quickly and accurately as possible, if it had the same *meaning* as the one presented previously. Therefore, its physical appearance (uppercase or lowercase) must be ignored. Half of the trials requested a positive answer. There were 30 trials. The score was the mean reaction time (RT) for the correct answers only.

In the numerical speed task, one single digit was displayed for 650 ms. After this presentation, a fixation point appeared for 500 ms. Finally, the probe digit appeared in order to decide, as quickly and accurately as possible, if it can be divided by the one presented previously. Half of the trials requested a positive answer. There were 30 trials. The score was the mean RT for the correct answers only.

In the spatial speed task, one arrow was displayed for 650 ms. The arrow can be displayed in one of seven orientations (multiples of 45°). After this presentation, a fixation point appeared for 500 ms. Finally, the probe arrow appeared in order to decide, as quickly and accurately as possible, if it had the same orientation of one presented previously. The arrows have distinguishable shapes in order to guarantee that their orientation is both memorized and evaluated. Half of the trials requested a positive answer. There were 30 trials. The score was the mean RT for the correct answers only.

Note that in these STM, WMC, and PS tasks, participants completed a set of three practice trials as many times as desired to ensure they understood the instructions.

Finally, attention (ATT) was measured by verbal and quantitative versions of the flanker task (Eriksen, & Eriksen, 1974) and a version of the Simon task (Simon, 1969). The verbal flanker task required deciding, as fast as possible, if the letter presented in the center of a set of three letters was vowel (by pressing the computer key 1) or consonant (by pressing the computer key 0). The target letter (e.g. vowel) can be surrounded by compatible (e.g. vowel) or incompatible (e.g. consonant) letters. The quantitative task required deciding, as fast as possible, if the digit presented in the center of a set of three digits was odd (by pressing the computer key 1) or even (by pressing the computer key 0). The target digit (e.g. odd) can be surrounded by compatible (e.g. odd) or incompatible (e.g. even) digits. The spatial task required deciding if an arrow (horizontally depicted) pointed to the left (by pressing the computer key 1) or to the right (by pressing the computer key 0) of a fixation point. The target arrow pointing to a given direction (e.g. to the left) can be presented at the left (e.g. compatible) or at the right (e.g. incompatible) of the fixation point. In all these tasks, there were a total of 32 practice trials and 80 experimental trials. Half of the trials were compatible and they were randomly presented across the entire session. The mean reaction time for the incompatible trials was the dependent measure.

2.3. Testing sessions (procedure)

Table 1 shows the organization of the successive testing sessions.

On week 1, participants were tested on the intelligence tests using the odd numbered items only. On weeks 2, 3, and 4, the STM-WMC group completed the three verbal, numerical, and spatial short-term memory and the three verbal, numerical, and spatial working memory tasks, whereas the PS-ATT group completed the three verbal, numerical, and spatial processing speed and the three verbal, numerical, and spatial attention tasks. Note that participants completed six different tasks on each session, so they really faced a total of 18 sub-sessions. As can be seen in Table 1, order of administration was counterbalanced across weeks. Finally, on week 5 participants were tested on the intelligence tests using the even numbered items only. Note that participants completed the five sessions exactly the same day at the same time every week. Finally, it is important to keep in mind that difficulty levels on the STM and WMC tasks were randomly delivered within and across sessions. Therefore, participants cannot predict the complexity level of the next trial from the previously presented, thus requiring systematic effortful processing.

Table 1
Testing sessions across weeks.

Week 1	Week 2	Week 3	Week 4	Week 5
Intelligence testing (odd items)		STM-WMC group		Intelligence retesting (even items)
	FDSPAN Reading span	FLSPAN Dot matrix	Corsi block Computation span	
	Corsi block Computation span	FDSPAN Reading span	FLSPAN Dot matrix	
DAT-SR	FLSPAN Dot matrix	Corsi block Computation span	FDSPAN Reading span	DAT-AR
APM		PS-ATT group		DAT-VR
	Verbal speed	Numerical speed	Spatial speed	
DAT-VR	Numerical attention	Spatial attention	Verbal atttention	DAT-SR
DAT-AR	Spatial speed	Verbal speed	Numerical speed	APM
	Verbal attention	Numerical attention	Spatial attention	
	Numerical speed	Spatial speed	Verbal speed	
	Spatial attention	Verbal attention	Numerical attention	

3. Results

Table 2 shows the descriptive statistics for both groups on the intelligence testing and retesting sessions.

Overall intelligence performance (APM + DAT-AR + DAT-VR + DAT-SR) increased from 47.2 at testing (SD = 10.5) to 52.9 (SD = 11.6) at retesting in the STM-WMC group (d = 0.51) and from 46.0 (SD = 8.6) to 52.3 (SD = 9.3) in the PS-ATT group (d = 0.70). Both groups increased their performance on the specific intelligence measures from testing to retesting, except on the DAT-AR test. The amount of improvement for APM, DAT-VR, and DAT-SR ranged from .46 to .73 d units. Performance differences between groups were negligible in both sessions, but specially at retesting (range from 0 to 0.15 d units).

The Appendix (A.1) shows the stability coefficients (test–retest) for the intelligence measures. The general score (APM +

Please cite this article as: Colom, R., et al., Improvement in working memory is not related to increased intelligence scores, *Intelligence* (2010), doi:10.1016/j.intell.2010.06.008

Table 2
Descriptive statistics for intelligence testing and retesting. Effect sizes (*d*) within (columns) and between (rows) groups are also shown.

		STM-WMC group (*N* = 173)		PS-ATT group (*N* = 115)		
		Mean	SD	Mean	SD	*d*
APM	Testing	10.61	2.66	10.37	2.72	*0.09*
	Retesting	11.93	3.1	12.10	2.9	*0.06*
	d	*0.46*		*0.61*		
DAT-AR	Testing	12.15	3.8	12.00	3.2	*0.04*
	Retesting	12.2	3.4	12.2	3.3	*0*
	d	*0.01*		*0.06*		
DAT-VR	Testing	11.7	2.9	11.9	2.7	*−0.07*
	Retesting	13.4	3.2	13.4	2.6	*0*
	d	*0.56*		*0.57*		
DAT-SR	Testing	12.7	4.2	11.7	3.7	*0.25*
	Retesting	15.3	5.1	14.6	4.2	*0.15*
	d	*0.56*		*0.73*		

DAT-AR + DAT-VR + DAT-SR) is highly stable (*r* = .83), whereas stability for the specific tests ranges from a minimum of .49 (DAT-VR) to a maximum of .68 (DAT-SR). The relatively low stability values for the specific measures could be attributed to subtests' length.

Table 3 shows the descriptive statistics for the STM-WMC group across sessions.

Improvements are evident for all the STM and WMC tasks. It is interesting to note, however, that the rate of improvement is generally much greater from first to the second session than from the second to the third session.

The Appendix (A.2) presents the stability coefficients for STM, WMC, as well as for a combination of both measures. It can be seen there is a high degree of stability in performance, with *r* values ranging from a minimum of .75 to a maximum of .88.

Table 4 shows the descriptive statistics for the PS-ATT group across sessions.

Table 3
Descriptive statistics for the STM-WMC group (*N* = 173) across the three testing sessions. Effect sizes (*d*) are also shown.

	Short-term memory			Working memory		
	FLSPAN	FDSPAN	Corsi block	Reading span	Computation span	Dot matrix
Session 1						
Mean	73.5	78.0	69.6	46.3	53.9	52.4
SD	14.8	13.2	13.7	16.3	14.7	6.3
Session 2						
Mean	79.4	86.2	73.7	48.5	57.9	55.7
SD	14.4	15.0	16.00	16.4	14.3	6.0
Session 3						
Mean	82.0	89.0	74.1	48.5	61.6	57.0
SD	16.11	15.7	16.3	17.7	12.2	6.9
Effect size						
S1–S2	*0.40*	*0.58*	*0.27*	*0.13*	*0.28*	*0.54*
S2–S3	*0.17*	*0.18*	*0.02*	*0*	*0.28*	*0.20*
S1–S3	*0.55*	*0.76*	*0.30*	*0.13*	*0.57*	*0.69*

Here the pattern of change is sharply different to that observed for the STM-WMC group. Now the rate of improvement (smaller RTs) generally remains or even increases across sessions.

The Appendix (A.3) presents the stability coefficients for PS and ATT, as well as for a combination of both measures across sessions. Although a bit lower than for STM and WMC, all the RT measures show high stability with *r* values ranging from a minimum of .69 to a maximum of .80.

Fig. 1. summarizes average changes (*d* units) for STM, WMC, PS and ATT from session 1 to session 2, as well as from session 2 to session 3. There is a great reduction of improvements for STM and WMC in session 3, while improvements remain at high levels for PS and ATT. We computed confidence intervals for the values shown in Fig. 1: STM (S1–S2 = .47 and .57; S2–S3 = .09 and .21), WMC (S1–S2 = .27 and .36; S2–S3 = .13 and .22), Speed (S1–S2 = .53 and .57; S2–S3 = .42 and .46), Attention (S1–S2 = .71 and .73; S2–S3 = .56 and .62). Therefore, improvements for STM and WMC in the last session are very small, whereas for Speed and Attention these improvements are still large.

We report descriptive statistics separated for males and females on summary measures for STM, WMC, Speed, and Attention across sessions, along with their intelligence scores on the pre and post-test sessions (see Appendix A). For the STM-WMC group, the male advantage increases across sessions: (a) from .30 to .37 on intelligence testing, (b) from .17 to .28 on STM, and (c) from .17 to .39 on WMC. For the Speed-ATT group, the male advantage on the first session vanishes and it becomes a female advantage on the second and third sessions both for speed and for attention. Intelligence performance is null on pretesting and there is a small male advantage on post-testing (.15).

Finally, Fig. 2 depicts correlations between average intelligence in the pretesting session and performance across sessions for STM, WMC, PS, and ATT. Interestingly, these correlations do not change across sessions for STM and WMC, whereas they show a systematic decrease for PS and ATT (*z* values for PS and ATT were 2.91 and 2.93 respectively, $p < .01$).

4. Discussion

4.1. Summary of findings

Here we have shown systematic improvements in intelligence and cognitive performance across several testing sessions. Increased intelligence scores were equivalent, on average, to more than half a standard deviation. Contrary to the main prediction, participants facing challenging and diverse (verbal, numerical, and spatial) memory span (short-term memory and working memory) task requiring high levels of effortful (complex) processing did not show better intelligence scores on post-testing, with respect to their intelligence scores in the pre-test, than participants facing simple and diverse (verbal, numerical, and spatial) processing speed tasks. Actually, the second group increased its intelligence performance to an equivalent of about 10 IQ points, whereas the first group increased its intelligence performance to an equivalent of about 7.5 IQ points.

It is very important to underscore that this general finding should not be directly compared with previous studies

Please cite this article as: Colom, R., et al., Improvement in working memory is not related to increased intelligence scores, *Intelligence* (2010), doi:10.1016/j.intell.2010.06.008

Table 4
Descriptive statistics for the PS-ATT group ($N=115$) across the three testing sessions. Effect sizes (d) are also shown.

	Processing speed			Attention		
	$\text{Verbal}_{\text{RT}}$	$\text{Numerical}_{\text{RT}}$	$\text{Spatial}_{\text{RT}}$	$\text{Verbal}_{\text{ATT}}$	$\text{Numerical}_{\text{ATT}}$	$\text{Spatial}_{\text{ATT}}$
Session 1						
Mean	545.4	885.3	583.7	568.0	636.6	478.1
SD	129.5	247.8	154.8	87.5	91.1	42.2
Session 2						
Mean	506.5	771.4	526.7	536.6	600.1	473.0
SD	118.3	257.7	130.4	77.4	89.6	60.4
Session 3						
Mean	469.2	687.5	483.5	504.4	561.6	440.4
SD	114.5	192.3	206.1	64.9	68.8	49.9
Effect size						
S1–S2	*0.31*	*0.45*	*0.40*	*0.38*	*0.41*	*0.10*
S2–S3	*0.32*	*0.37*	*0.25*	*0.46*	*0.47*	*0.60*
S1–S3	*0.62*	*0.89*	*0.55*	*0.83*	*0.93*	*0.82*

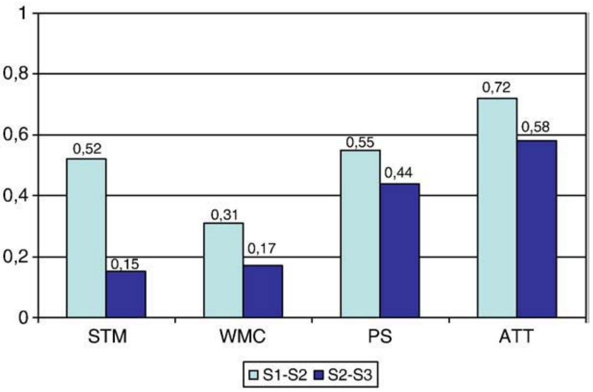

Fig. 1. Improvements in performance (d units) from session 1 (S1) to session 2 (S2), as well as from session 2 (S2) to session 3 (S3). STM = short-term memory, WMC = working memory, PS = processing speed, ATT = attention.

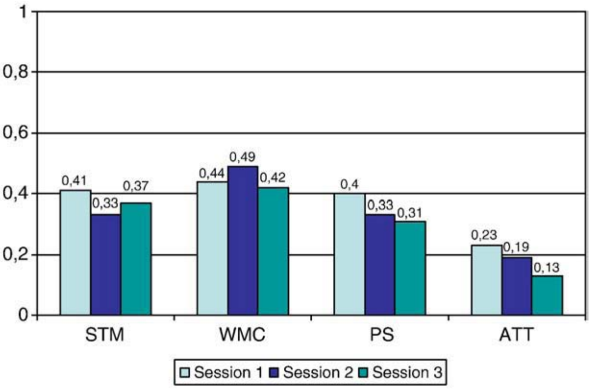

Fig. 2. Correlations between intelligence at pretesting and average performance across sessions in STM (short-term memory), WMC (working memory), PS (processing speed), and ATT (attention).

analyzing the presumed effects of cognitive 'training', such those reported by Jaeggi et al. (2008) or the research summarized by Buschkuehl et al. (2008). Training per se is not the main issue here. We were looking for predicted differential improvements on intelligence scores after repetitive exposure to cognitively challenging requirements vs. repetitive exposure to simple speed tasks. As noted by one reviewer of the present paper, it would be really interesting to analyze two further groups: (a) a simple pretest-posttest with no training to measure simple practice effects, and (b) a more challenging condition with longer duration. Furthermore, dose-response results cannot be analyzed here because the two groups were exposed to 18 different tasks across three sessions. We acknowledge that dose-response effects are probably relevant, but the applied design is not suited for testing this point.

Beyond the general finding, there are several noteworthy results. First, participants did show better intelligence scores in three out of four measures. The exception was the abstract reasoning subtest from the DAT battery (DAT-AR). To find out an explanation for this surprising result, we computed an exploratory factor analysis (Principal Axis Factoring) for both pre and post testing measures. Results revealed that the abstract reasoning test shows the highest loading on the extracted factor (.76 and .82 in the testing and retesting sessions, respectively). Loadings for the remaining intelligence tests ranged from .51 (DAT-VR) to .63 (APM) in the testing session and from .56 (DAT-VR) to .66 (APM) in the retesting session. Therefore, accepting that the obtained factor represents g, it might be concluded that participants did not show any change in their performance on the most g loaded intelligence measure (Jensen, 1998)[1].

[1] One reviewer suggested that it could be that the even items are simply more difficult than the odd items for some reason. However, we computed a t test after the difficulty levels for even (mean = .600) and odd (mean = .599) DAT-AR items and we failed to find a significant difference ($p=.967$).

Nevertheless, a systematic effect is required for supporting this conclusion, as noted by one reviewer. For illustrative purposes only, the *g* loadings after the four intelligence subtests were obtained from the two groups. Afterwards, Pearson (*r*) and Spearman (Rho) correlations between *g* loadings and pretest–posttest changes (*d* units from Table 2) were computed. For the STM-WMC group the values were $r = -.83$ and Rho = −.99, whereas for the Speed-ATT group the values were $r = -.40$ and Rho = −.78. Thus, there is a trend showing that the greater the *g* loading the smaller the change from the pre to the posttest intelligence session.

Second, increased performance in STM and WMC was relevant from the first to the second session (equivalent to 8 and 5 IQ points, respectively), but it was much smaller from the second to the third session (equivalent to 2 IQ points). This contrast with the rate of improvement for PS and ATT: from the first to the second session the change was equivalent to 8 and 11 IQ points respectively, whereas from the second to the third session the change was equivalent to 7 and 9 IQ points respectively. These results suggest that participants approached their asymptotic STM and WMC level on the third session, while this was not the case for PS and ATT.

Third, the predictive validity for intelligence at pretesting remains across STM and WMC sessions: for STM *r* values were .41 (first session) and .37 (third session) whereas for WMC *r* values were .44 (first session) and .42 (third session). However, this predictive validity shows a monotonic decrease across sessions for PS and ATT, with *r* values falling from .40 (first session) to .13 (third session). Therefore, even when participants still show large improvements in their PS and ATT performance on session 3, these improvements appear to be independent of participants' intelligence. This is just the opposite of what we have seen for STM and WMC: participants show meager improvements on the third session, but their performance still relates to intelligence scores in the pre-testing session.

Finally, the relevant changes observed across sessions do not impact on the stability of performance. For overall intelligence, we reported a stability coefficient of .83. This coefficient was also very high for the analyzed cognitive functions, with values ranging from a minimum of .75 to a maximum of .88. Actually, the correlation for STM + WMC between the first and the third session was .80, and the correlation for PS + ATT between the first and the third session was a bit lower but still large (.75). Therefore, we can conclude that there is a high degree of stability in cognition, no matter the ability (general intelligence, short-term storage, working memory capacity, processing speed, or attention) you are tapping.

4.2. Why should improvements in working memory be related to increased intelligence scores?

The seminal paper by Kyllonen, and Christal (1990) proposed that working memory capacity (WMC) is central for the human information processing system. Further, reasoning ability is central for conventional ability models (Carroll, 1993;2003; Johnson, & Bouchard, 2005). They investigated the degree of correlation between theses psychological constructs, finding a result that opened the door to their main conclusion: reasoning ability is little more than working memory capacity.

Kyllonen and Christal suggested that working memory capacity is responsible for differences in reasoning ability, because WMC mental processes are central to the component stages of reasoning tasks. However, they recognised that WMC could be primarily determined by individual differences in reasoning ability: "*part of WMC might be simply characterized as the size of storage buffers. But the analysis here could be interpreted as suggesting that an additional important determinant of memory capacity is the degree to which buffer storage can be managed through a kind of reasoning process. To be successful in WMC tasks requires the ability to reason successfully about how to manage short-term storage resources*" (p. 428).

Colom, Abad, Rebollo, and Shih (2005);Colom et al. (2004) replicated the very high relationship between WMC and intelligence across samples from different countries and different test batteries, concluding, like Kyllonen and Christal, that these constructs are almost isomorphic. This general conclusion remains unchallenged, when, and only when, the constructs are appropriately tapped. Nevertheless, there are some researchers claiming that the correlation is not that large (Ackerman, Beier, & Boyle, 2005, but see Kane, Hambrick, & Conway, 2005; Oberauer, Schulze, Wilhelm, & Süb, 2005).

As referred at the introduction section, Halford et al. (2007) proposed that the large correlation between WMC and intelligence reflects a common capacity limitation: (a) WMC temporarily binds elements to a system associated to relational representations necessary for reasoning, (b) the shared capacity limit depends on the number of bindings (approx 4 on average), and (c) WMC comprises a domain-general component accounting for the correlation between WMC and intelligence. Consistent with this approach, Colom et al. (2008) found, across three related studies using progressively more relevant measures (short-term storage, processing speed, updating executive processes, and attention) that the high relationship between WMC and intelligence can be accounted for by simple short-term storage (STM). In their last study comprising all the relevant constructs, STM was the main predictor of intelligence.

In addition to these behavioural approaches, neuroimaging evidence is also consistent with the view that WMC and intelligence share relevant neural mechanisms (Gray, Chabris, & Braver, 2003; Kane, & Engle, 2002; Marois, & Ivanoff, 2005). Using a VBM approach, Colom, Jung, and Haier (2007) found that there is a common neuroanatomic base for memory span and the general factor of intelligence (*g*) relying on a fronto-parietal network (Jung, & Haier, 2007). The lesion study reported by Gläscher et al. (2010) also supports the relevance of this biological common substrate.

From this cumulated behavioural and biological evidence, it seems reasonable to predict that improvements in WMC should differentially impact on intelligence performance. This prediction was behind the Jaeggi et al. (2008) study. These researchers reported a positive effect, but their results would be seen with reservations (Moody, 2009). The study reported here did not find the predicted differential improvement. Even when we have seen relevant increased scores from a pre-test to a post-test intelligence session separated by one month, we did not find any differential improvement between one group repeatedly facing highly demanding (complex) processing and one group repeatedly facing cognitive tasks based on simple speeded performance. The fact that (a) the rate of performance improvements in tasks requiring high degrees of effortful (complex) processing was much smaller across sessions, but still related to pre-test intelligence scores, whereas (b) the rate of performance improvements in tasks requiring very low levels of effortful (simple) processing remained high across sessions, but progressively least correlated with pre-test intelligence scores, is consistent with the main 'manipulation' attempted in the present study. Further, even when the evidence is weak, we have shown that participants show less improvements in more highly *g* loaded intelligence measures. This result supports the distinction, underscored by Jensen (1998) among constructs (*g*), vehicles (psychometric tests or cognitive tasks), and measurements (accuracy scores or reaction times).

Appendix A

A.1. Stability coefficients for the intelligence measures ($N = 288$). All these correlations are significant at $p < .01$. Values with bold emphases highlight the stability coefficients.

	1	2	3	4	5	6	7	8	9	10
1. APM-PRE		.48	.31	.39	***.52***	.55	.41	.43	.70	.62
2. DAT-AR-PRE			.39	.46	.39	***.61***	.39	.49	.80	.62
3. DAT-VR-PRE				.31	.28	.43	***.49***	.36	.65	.51
4. DAT-SR-PRE					.40	.51	.37	***.68***	.78	.67
5. APM-POST						.53	.42	.37	.53	.72
6. DAT-AR-POST							.44	.54	.71	.82
7. DAT-VR-POST								.32	.56	.67
8. DAT-SR-POST									.68	.80
9. General PRE										***.83***
10. General POST										

A.2. Stability coefficients for the STM and WMC measures ($N = 173$).

	1	2	3	4	5	6	7	8	9
1. STM-S1		***.79***	***.75***	.60	.56	.53	.90	.77	.73
2. STM-S2			***.85***	.50	.58	.52	.72	.91	.78
3. STM-S3				.53	.60	.63	.72	.83	.93
4. WMC-S1					***.83***	***.77***	.89	.73	.70
5. WMC-S2						***.86***	.77	.86	.79
6. WMC-S3							.72	.76	.87
7. Session 1								***.84***	***.80***
8. Session 2									***.88***
9. Session 3									

A.3. Stability coefficients for the PS and ATT measures ($N = 115$).

	1	2	3	4	5	6	7	8	9
1. PS-S1		***.69***	***.69***	.52	.51	.50	.95	.68	.68
2. PS-S2			***.73***	.56	.73	.67	.73	.97	.76
3. PS-S3				.55	.66	.72	.73	.75	.98
4. ATT-S1					***.71***	***.71***	.77	.65	.63
5. ATT-S2						***.77***	.66	.87	.73
6. ATT-S3							.65	.74	.85
7. Session 1								***.75***	***.75***
8. Session 2									***.80***
9. Session 3									

A.4. Means and standard deviations (SD) for males and females from the STM-WMC ($N = 39$ males and 134 females) and Speed-ATT ($N = 13$ males and 102 females) groups on summary scores for STM, WMC, Speed, and Attention across sessions (S1, S2, and S3) along with their intelligence scores on the pre and post-test sessions. The average difference in *d* units is also shown.

	Males		Females		*d*
	Mean	SD	Mean	SD	
STM-WMC group					
Pre Intelligence	49.7	9.8	46.6	10.5	.30
STM-S1	225.5	35.1	219.9	31.6	.17
STM-S2	246.4	36.3	237.3	36.8	.25
STM-S3	253.5	36.0	242.6	40.2	.28
WMC-S1	156.8	28.1	151.5	30.8	.17
WMC-S2	169.7	26.8	159.8	30.3	.33
WMC-S3	176.3	27.8	164.6	30.5	.39
Post Intelligence	56.2	11.6	51.9	11.6	.37
Speed-ATT group					
Pre Intelligence	46.00	6.5	45.9	8.8	.01
Speed-S1	649.2	148.5	669.5	137.8	−.15
Speed-S2	606.0	134.6	592.7	126.6	.10
Speed-S3	562.0	136.2	536.4	121.0	.21
ATT-S1	543.5	77.5	560.7	70.2	−.24
ATT-S2	529.1	72.7	509.9	56.5	.33
ATT -S3	486.7	63.1	480.2	45.4	.13
Post Intelligence	53.6	6.9	52.2	9.6	.15

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*, 30−60.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1990). *Differential Aptitude Test (5th Edition)*. Madrid: TEA.

Brody, N. (1992). *Intelligence (2nd ed.)*. San Diego, CA: Academic Press.

Brown, A. L., & Campione, J. C. (1982). Modifying intelligence or modifying cognitive skills: More than a semantic quibble? In D. K. Detterman, & R. J. Sternberg (Eds.), *How and how much can intelligence be increased* (pp. 215−230). NJ, Ablex: Norwood.

Buschkuehl, M., Jaeggi, S. M., Hutchison, S., Perrig-Chiello, P., Däpp, Ch., Müller, M., et al. (2008). Impact of working memory training on memory performance in old–old adults. *Psychology and Aging*, *23*(4), 743−753.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404−431.

Carroll, J. B. (1993). *Human Cognitive Abilities*. Cambridge: Cambridge University Press.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about 10 broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen*. (pp. 5−21)Amsterdam: Pergamon.

Colom, R., Abad, F. J., Quiroga, Mª. A., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence*, *36*, 584−606.

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

Colom, R., Jung, R. E., & Haier, R. J. (2007). General intelligence and memory span: Evidence for a common neuro-anatomic framework. *Cognitive Neurosychology*, *24*, 867−878.

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

Daneman, M., & Carpenter, P. A. (1980). Individual-differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450−466.

Elosúa, M. R., Gutiérrez, F., García-Madruga, J. A., Luque, J. L., & Gárate, M. (1996). Adaptación española del "Reading Span Test" de Daneman y Carpenter [Spanish standardization of the Reading Span Test]. *Psicothema*, *8*, 383−395.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of target letter in a non-search task. *Perception & Psychophysics*, *16*, 143—149.

Gläscher, J., Rudrauf, D., Colom, R., Paul, L. K., Tranel, D., Damasio, H., & Adolphs, R. (2010). The distributed neural system for general intelligence revealed by lesion mapping. *PNAS*. www.pnas.org/cgi/doi/10.1073/pnas.0910397107.

Gottfredson, L., et al. (1997). Mainstream science on Intelligence: An Editorial with 52 signatories, history, and bibliography. *Intelligence*, *24* (1), 13—23.

Gray, J., Chabris, C., & Braver, T. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, *6*, 316—322.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences*, *11*, 236—242.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *PNAS*, *105*, 6829—6833.

Jensen, A. R. (1998). *The g factor*. Westport, Connecticut, Praeger: The science of mental ability.

Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*, 393—416.

Jung, R. E., & Haier, R. J. (2007). The Parieto-Frontal Integration Theory (P-FIT) of Intelligence: Converging Neuroimaging Evidence. *The Behavioral and Brain Sciences*, *30*, 135—187.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*, 637—671.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 66—71.

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

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, *9*, 296—305.

Moody, D. E. (2009). Can intelligence be increased by training on a task of working memory? *Intelligence*, *37*, 327—328.

Neisser, U., Boodoo, G., Bouchard, T., Boykin, A., Brody, N., Ceci, S., et al. (1996). Intelligence: Knowns and unknowns. *The American Psychologist*, *51*(2), 77—101.

Oberauer, K., Schulze, R., Wilhelm, O., & Süb, H. (2005). Working memory and intelligence—Their correlation and their relation: Comment on Ackerman, Beier, and Boyle. *Psychological Bulletin*, *131*, 61—65.

Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology*, *81*, 174—176.