Intelligence 41 (2013) 712–727

Contents lists available at ScienceDirect

Intelligence

Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing

Roberto Colom [a,*], Francisco J. Román [a], Francisco J. Abad [a], Pei Chun Shih [a], Jesús Privado [b], Manuel Froufe [a], Sergio Escorial [b], Kenia Martínez [a], Miguel Burgaleta [a,c], M.A. Quiroga [b], Sherif Karama [d], Richard J. Haier [e], Paul M. Thompson [f,g], Susanne M. Jaeggi [e]

[a] Universidad Autónoma de Madrid, Spain
[b] Universidad Complutense de Madrid, Spain
[c] Universidad Pompeu Fabra, Barcelona, Spain
[d] Montreal Neurological Institute (MNI), Canada
[e] University of California at Irvine (UCI), United States
[f] University of California at Los Angeles (UCLA), United States
[g] University of Southern California (USC), United States

ARTICLE INFO

Article history:
Received 22 May 2013
Received in revised form 6 September 2013
Accepted 6 September 2013
Available online xxxx

Keywords:
Cognitive training
Intelligence
Working memory
Attention

ABSTRACT

Short-term adaptive cognitive training based on the n-back task is reported to increase scores on individual ability tests, but the key question of whether such increases generalize to the intelligence construct is not clear. Here we evaluate fluid/abstract intelligence (Gf), crystallized/verbal intelligence (Gc), working memory capacity (WMC), and attention control (ATT) using diverse measures, with equivalent versions, for estimating any changes at the construct level after training. Beginning with a sample of 169 participants, two groups of twenty-eight women each were selected and matched for their general cognitive ability scores and demographic variables. Under strict supervision in the laboratory, the training group completed an intensive adaptive training program based on the n-back task (visual, auditory, and dual versions) across twenty-four sessions distributed over twelve weeks. Results showed that this group had the expected systematic improvements in n-back performance over time; this performance systematically correlated across sessions with Gf, Gc, and WMC, but not with ATT. However, the main finding showed no significant changes in the assessed psychological constructs for the training group as compared with the control group. Nevertheless, post-hoc analyses suggested that specific tests and tasks tapping visuospatial processing might be sensitive to training.

1. Introduction

General intelligence (*g*) is defined by a broad ability for reasoning, solving problems, and efficient learning (Gottfredson et al., 1997). Hunt (1995, 2011) underscores the distinction between fluid intelligence (Gf) and crystallized intelligence (Gc), although these two broad abilities are related with *g* (Carroll, 1993, 2003; McGrew, 2009). Gc involves the intelligent use of culturally rooted knowledge and skills (such as language or math), whereas Gf requires abilities for solving novel and abstract problems (Cattell, 1987). These latter abilities are the main target of cognitive training programs.

It has been repeatedly demonstrated that tests' scores can be increased (Neisser et al., 1996; Nisbett et al., 2012) but, is it possible to improve cognitive ability? (Colom et al., 2010). For obtaining convincing evidence, the training tools must be substantially different to usual tests of cognitive ability. Test specific skills can be improved by increased familiarity

[*] Corresponding author at: 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 © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.intell.2013.09.002

(te Nijenhuis, van Vianen, & van der Flier, 2007) but this is hardly interesting. There are reports showing improvements in intelligence, as assessed by standard tests, after training information processing skills. Thus, for instance, Posner and Rothbart (2007) trained children on a visual attention task based on the management of conflict. The trained children scored higher than a control group on a standard intelligence battery. Also studying children, Irwing, Hamza, Khaleefa, and Lynn (2008) reported large improvements in the Raven Progressive Matrices Test in a group trained for several months with the abacus (requiring the reliable preservation of intermediate calculations in working memory) when compared with a non-trained group. Jaušovec and Jaušovec (2012) reported a positive effect in the RAPM test after working memory training; the change from the pretest to the posttest assessment was equivalent to thirteen IQ points ($d = .88$) for their training group, whereas it was null for an active control group. Digit span scores were also substantially higher for the training group ($d = 0.81$) than for the control group ($d = 0.25$). The study by von Bastian and Oberauer (2013) concluded that general reasoning ability can be improved by working memory training (self-administered at home). The positive effect was also observed six months after ending the training program. Further, training of specific working memory processes (storage and processing, relational integration, or supervision) led to transfer in specific cognitive factors.

Jaeggi, Buschkuehl, Jonides, and Perrig (2008) reported that training in a challenging adaptive dual n-back task (tapping a mixture of executive updating, working memory, and attention skills) was related to better performance on a fluid intelligence test compared to a passive control group. These results were repeated in further studies: (a) training on the single n-back task (either visual or verbal) showed similar positive effects over performance on fluid intelligence tests (Jaeggi, Buschkuehl, Shah, & Jonides, in press; Jaeggi et al., 2010) and (b) similar findings were observed in a sample of children (Jaeggi, Buschkuehl, Jonides, & Shah, 2011). However, the conclusion that n-back training improves fluid intelligence is controversial. For instance, Moody (2009) argued that improvements on the specific fluid measure considered by Jaeggi et al. (2008) could be explained by the strict time limit imposed for solving the less difficult items. In his view, no challenge was made over the participants' Gf, and, therefore, observed changes may be fully explained by a speed factor. From a broader perspective, Shipstead, Redick, and Engle (2012) argued that these short-term training studies fail to really increase abilities required by the working memory processing system. In their view, published studies (1) generally rely on single measures for measuring predicted intelligence changes after training and (2) administer invalid measures of working memory capacity. These authors note that a wide variety of tasks measuring the constructs of interest must be systematically administered in order to avoid critiques related to task specificity issues.

A study by Chooi and Thompson (2012) was aimed at overcoming some of the reservations enumerated by Shipstead et al. (2012) and used several measures of intelligence (crystallized-verbal, spatial, and speed) for estimating changes after training on the dual n-back task modeled from Jaeggi et al. (2008). Working memory was measured by a single task (operation span). They failed to find any effect of training on either intelligence or working memory. Passive and active controls were considered along with the training group, using two time lengths (8 and 20 days) resulting in very small sample sizes for the six analyzed groups (from 9 to 23 participants). Importantly, the n-back performance level achieved by the trained participants in the 20-day training period was well below the one attained by the Jaeggi et al.'s sample.

Redick et al. (2012) reported a similar study. Fluid (six tests) intelligence and crystallized (two tests) intelligence, along with working memory (two tasks), multitasking (three tasks), and processing speed (two tasks) were the measured constructs. Training ($N = 24$), active ($N = 29$), and passive ($N = 20$) control groups were analyzed. Very short versions of the considered psychological tests were administered before, during, and after the training stage. This study failed to find any difference among the three groups, consistent with Chooi and Thompson (2012). Surprisingly, Redick et al. failed to find any practice effect across their three evaluations. Again, n-back performance level achieved by the trained participants was almost identical to the attained in Chooi and Thompson and well below the reported by Jaeggi et al. (2008, 2010).

Recently, Stephenson and Halpern (2013) replicated the Jaeggi et al.'s (2008, 2010) main findings. However, significant gains were observed in two out of four fluid intelligence tests (RAPM and BETA-Matrix Reasoning). Thus, for instance, after the adaptive training program based on the dual n-back ($N = 28$) gains were equivalent to (a) 13.3 IQ points ($d = 0.89$) in the BETA-Matrix Reasoning (b) 9.9 IQ points ($d = 0.66$) in the RAPM, (c) 8.4 IQ points ($d = 0.56$) in the WASI-Matrix Reasoning, and (d) 5.2 IQ points ($d = 0.35$) in the Culture-Fair Intelligence Test.

The theoretical framework for the present study is based on the available evidence demonstrating a very high correlation between intelligence and working memory at the latent variable level (Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004; Oberauer, Schulze, Wilhelm, & Süb, 2005). The comprehensive study by Martínez et al. (2011) is a recent example considering twenty-four measures tapping eight intelligence and cognitive factors (three measures for each factor): fluid-abstract intelligence, crystallized-verbal intelligence, and spatial intelligence, along with short-term memory, working memory capacity, executive updating, attention, and processing speed. Their main findings support the view that fluid intelligence can be largely identified with basic short-term storage processes tapped by working memory tasks and executive updating. This was seen as quite consistent with neuroimaging results showing that fluid intelligence shares relevant brain structural (Colom, Jung, & Haier, 2007) and functional (Gray, Chabris, & Braver, 2003) correlates with working memory capacity. The large correlation between intelligence and working memory at the latent variable level suggests that they share substantial capacity limitations based on the amount of information that can be reliably kept active in the short-term, both within the working memory system or during the reasoning processes required on intelligence tests (Colom, Abad, Quiroga, Shih, & Flores-Mendoza, 2008; Colom, Rebollo, Abad, & Shih, 2006; Halford, Cowan, & Andrews, 2007).

The proper testing of the prediction that improvements in the working memory system (short-term storage and executive updating) through adaptive cognitive training will promote increments in fluid intelligence, mainly because their common limitations for the reliable temporary storage of the relevant information will be boosted, requires straightforward analyses going well beyond the level of specific measures, as underscored by Shipstead et al. (2012). For that purpose, here we administered several diverse intelligence and cognitive measures (three measures for each psychological factor) before and after completing a challenging cognitive training program based on the adaptive n-back dual task firstly proposed by Jaeggi et al. (2008).

The main prediction is that if adaptive working memory training promotes skills relevant for the reliable temporary storage of relevant information, then fluid intelligence and working memory scores will be higher for the trained than for the control group at the posttest evaluation. Further, (a) these higher scores must be systematically observed for all Gf and WMC specific tests and tasks, and (b) crystallized intelligence and attention control will not be sensitive to training. Both fluid intelligence and working memory require the reliable preservation of the relevant information in the short-term, as demonstrated by the seminal study by Carpenter, Just, and Shell (1990). This is not the case for crystallized intelligence and attention control, because Gc requires the recovery of the relevant information from long-term memory and attention control does not requires any short-term storage.

2. Method

2.1. Participants

One hundred and sixty nine psychology undergraduates completed a battery of twelve intelligence tests and cognitive tasks measuring fluid-abstract intelligence, crystallized-verbal intelligence, working memory capacity, and attention control. After computing a general index from the six intelligence tests, two groups of twenty-eight females were recruited for the study. They were paid for their participation.[1] Members of each group were carefully matched for their general intelligence index, so they were perfectly overlapped and represented a wide range of scores. All participants were right handed, as assessed by the Edinburgh Test (Oldfield, 1971). They also completed a set of questions asking for medical or psychiatric disorders, as well as substance intake. The recruitment process followed the Helsinki guidelines (World Medical Association, 2008) and the local ethics committee approved the study. Descriptive statistics for the demographic variables and performance on the cognitive measures for the two groups of participants (training and control) can be seen in Appendix A (Table A.1).

2.2. Basic design

The collective psychological assessment for the pretest stage was done from September 19 to October 14, 2011. Participants were assessed in groups not greater than twenty-five. The data obtained for the complete group ($N = 169$) were analyzed for recruiting the training ($N = 28$) and control ($N = 28$) groups based on the general index computed from the measures of fluid intelligence and crystallized intelligence (Table A.1). The adaptive cognitive training program began in November 14, 2011, remained active until February 17, 2012, and lasted for twelve weeks (with a break from December 24, 2011 to January 9, 2012). The psychological assessment for the posttest was done individually from February 20 to March 09 (intelligence tests) and from March 12 to March 30 (cognitive tasks), 2012.

2.3. Psychological constructs

Intelligence and cognitive constructs were assessed by three measures each. As noted above, fluid intelligence (Gf) requires abstract problem solving abilities, whereas crystallized intelligence (Gc) involves the mental manipulation of cultural knowledge. Gf was measured by screening versions (odd numbered items and even numbered items for the pretest and posttest evaluations, respectively) of the Raven Advanced Progressive Matrices Test (RAPM), the abstract reasoning subtest from the Differential Aptitude Test (DAT-AR), and the inductive reasoning subtest from the Primary Mental Abilities Battery (PMA-R). Gc was measured by screening versions (odd numbered items and even numbered items for the pretest and posttest evaluations, respectively) of the verbal reasoning subtest from the DAT (DAT-VR), the numerical reasoning subtest from the DAT (DAT-NR), and the vocabulary subtest from the PMA (PMA-V). Gf and Gc were measured by tests with (PMA subtests) and without (RAPM and DAT subtests) highly speeded constraints. Working memory capacity requires the simultaneous processing and storage of varied amounts of information. WMC was measured by the reading span, the computation span, and the dot matrix tasks. Finally, attention control was tapped by cognitive tasks based on the quick management of conflict: verbal (vowel–consonant) and numerical (odd–even) flanker tasks, along with the spatial (right–left) Simon task. The working memory capacity and attention control tasks were the same for the pretest and the posttest sessions. A detailed description of these intelligence tests and cognitive tasks can be found in Appendix A (Table A.2). Fig. 1 shows examples of the intelligence and cognitive tasks.

2.4. Cognitive training schedule

The framework for the cognitive training program followed the guidelines reported by Jaeggi et al. (2008) but it was re-programmed for Visual Basic (2008 Version). Nevertheless, there were some differences: (a) the training began with four sessions (weeks 1 and 2) with a visual adaptive n-back version and four sessions (weeks 3 and 4) with an auditory adaptive n-back version before facing the sixteen sessions of the adaptive n-back dual program (weeks 5 to 12), and (b) while the training program is usually completed in one month, here we extended the training period to three months (12 weeks). There were two training sessions per week lasting around 30 min each and they took place under strict supervision in the laboratory. Participants worked within individual cabins and the experimenter was always available for attending any request they might have. Data were analyzed every week for checking their progress at both the individual and the group level. Participants received systematic feedback regarding their performance. Furthermore, every two weeks, participants completed a motivation questionnaire asking for their (a) involvement with the task, (b) perceived difficulty level, (c) perceived challenging of the task levels, and (d) expectations for future achievement. At the end of the training period, participants were asked with respect to

[1] 200 € if assigned to the training group and 100 € if assigned to the control group.

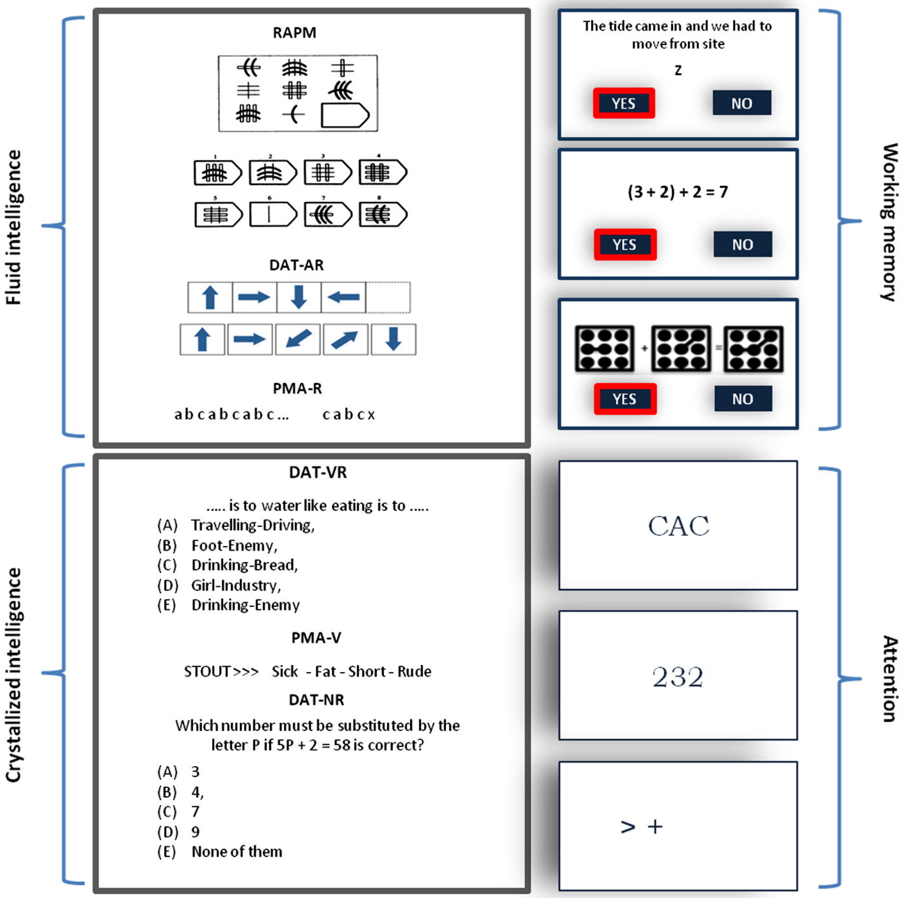

Fig. 1. Examples of intelligence items and cognitive tasks. Left panel shows example items for fluid intelligence (Gf) — Raven Advanced Progressive Matrices Test (RAPM), abstract reasoning (DAT-AR), and inductive reasoning (PMA-R), and crystallized intelligence (Gc) — verbal reasoning (DAT-VR), vocabulary (PMA-V), and numerical reasoning (DAT-NR). Right panel depicts examples for working memory capacity (reading span, computation span, and dot matrix) and attention control (vowel–consonant, odd–even, and right–left).

their general evaluation of the program. Using a rating scale from 0 to 10, average values were (a) 8.1 (range 8.0 to 8.2 across sessions), (b) 7.9 (range 7.4 to 8.5 across sessions), (c) 8.0 (range 7.8 to 8.2 across sessions), and (d) 7 (range 6.5 to 7.7 across sessions).

The control group was passive. After the recruitment process, members of this no-contact control group were invited to follow their normal life as university students. As reasoned in some of our previous research reports addressing the potential effect of cognitive training, and according to the main theoretical framework, we were not interested in comparing different types of training, but in the comparison between a specific cognitive training and doing nothing beyond regular life. The difference between passive vs. active controls is relevant when two different treatments are compared. If the issue is to compare participants doing physical exercise vs. not doing any exercise, it is uninteresting to compare jogging and body building, for example. Here, we are not contrasting the effect of theoretically different training programs, but training vs. not training (Colom et al., 2012; Martínez et al., in press). Further, (a) Chooi and Thompson (2012) and Redick et al. (2012) failed to find any difference between their active and passive control groups using a closely similar approach, and (b) the meta-analysis published by Klauer and Phye (2008) did not observe any difference between no-contact and placebo groups.

2.5. Analyses

First, the achieved average n-back level was computed for all visual, auditory, and dual training sessions. In addition, correlations between intelligence/cognitive pretest scores and individual differences in achieved n-back level across sessions were computed.

Second, pretest and posttest scores on the two intelligence factors (Gf and Gc) and the six intelligence tests were transformed using item response theory (IRT) for equating their level of difficulty, making them strictly comparable for the training and control groups. These IRT calculations were obtained from independent samples, as explained in full detail in Appendix A.3. Shortly, we followed three steps: (1) calibrating odd and even items in the same sample (for obtaining IRT odd and even item parameters using the same metric), (2) with item parameters fixed to those obtained in the previous phase, we applied IRTPRO separately to the odd test and the even test, obtaining two conversion tables (one for the odd part and one for the even part). Each conversion table indicated what θ corresponds to each sum score in this part. The same prior distribution for θ (m = 0; s = 1) was assumed when computing the conversion table, and (3) conversion tables were applied to the training and control groups for obtaining IRT scores from sum scores.

We applied IRT because: (a) IRT provides better scaling of individual differences relative to the raw score metric (Embretson & Reise, 2000; Reise & Haviland, 2005). Classic test theory assumes linear relationship between true and observed test scores, and also, that precision is equal across the ability range. IRT scaling takes into account non-linear relationships between traits and observed scores, and the known fact that the standard error differs across trait levels; (b) when different forms are applied in pretest and the posttest sessions, as in our case, IRT is an ideal tool for adjusting differences in difficulty and precision of the administered forms. Unlike classic test models, IRT can concurrently separate both the effect of examinee's ability and item characteristics (e.g., difficulty); (c) IRT has additional advantages for facilitating scores interpretation in terms of probability of successfully solve one specific task for one specific score.

Finally, the training and control groups were compared at the constructs and at the measures levels. The main goal here is the analysis at the construct level, but constructs are not homogeneous and, therefore, results at the measures level also deserve inspection. Standardized changes were computed after the following formula: (posttest − pretest) / $SD_{pretest}$. These standardized changes were submitted to analyses of covariance (ANCOVA) where the group was the independent variable, the construct/measure was the dependent variable, and the covariate was the score at the pretest for the corresponding variable. A p level of .05 (one-tailed) was considered for testing the results (Jaeggi et al., 2008). Note that a post-hoc power analysis (G*Power; Faul, Erdfelder, Lang, & Buchner, 2007) indicated that we had sufficient power to detect a significant Group (between-subjects) × Session (within-subjects) interaction, if it was present in the transfer data. The power to detect a large ($f = .40$) or medium ($f = .25$) effect was >.99, based on the sample size and the use of the within-subjects correlation of $r = .84$ (which was the largest correlation among the repeated measures across all 12 transfer tasks). We also re-ran the power analyses using the smallest correlation among repeated measures of $r = .14$ (Verbal Flankers). In this case, the power to detect a large or medium effect was >.80. For ANCOVA analyses, the power was .84 for detecting a large effect size ($f = .40$) and .45 for a medium ($f = .25$) effect size.

3. Results

Fig. 2 depicts results for the average n-back levels achieved by the training group across the visual, auditory, and dual sessions. Large improvements were found for the three versions. Indeed, the achieved final average level for the dual version (5.13) was almost identical to that reported by Jaeggi et al. (2008). These levels ranged from 3–4 to 9–10 back. Following Chooi and Thompson (2012), we also obtained the percentage of improvement for each condition (average achieved level in the last session minus average level in the first session). The result was divided by the level achieved in last session and multiplied by 100. For the visual condition the improvement was 41%, for the auditory condition it was 39%, and for the dual condition it was 53%.

We also computed the correlation between pretest intelligence/cognitive factors and achieved n-back levels across the full range of training sessions (Appendix A.4 (Fig. A.4)). Interestingly, (a) the correlations for fluid intelligence, crystallized intelligence, and working memory capacity were systematically statistically significant (above .40), whereas for attention control they were not significant across sessions, and (b) it is noteworthy that the correlation between fluid intelligence and achieved n-back level on the dual version increased across sessions, but this is not the case for crystallized intelligence and working memory capacity.

Results for the changes observed from the pretest to the posttest assessments will be firstly presented at the construct level. Nevertheless, changes at the test level will be also analyzed for providing a detailed picture of participants' performance. As noted above, the analysis at the construct level is the main goal of the present study, but constructs are heterogeneous and therefore the inspection of results for their specific measures may provide relevant knowledge.

3.1. Changes at the construct level

Fig. 3 depicts results for the considered constructs. Note that IRT transformations were the input data for fluid intelligence and crystallized intelligence (Appendix A.3). Fluid intelligence (Gf) increased for both groups from the pretest to the posttest; the effect size (d) was 0.81 for the training group and 0.46 for the control group. Crystallized intelligence (Gc) did not change from the pretest to the posttest in both groups; the effect size (d) was −0.03 for the training group and 0.07 for the control group. Working memory increased for both groups from the pretest to the posttest; the effect size (d) was 0.54 for the training group and 0.41 for the control group. Finally, attention control improvements for both groups from the pretest to the posttest were small; the effect size (d) was 0.26 for the training group and 0.12 for the control group.

Fig. 4 depicts the computed standardized changes for the results shown in Fig. 3. There are no significant differences between the training and control groups for any construct (except for fluid intelligence at a trend level, $p = .06$). Therefore, the noted improvements of the training group in the adaptive n-back program (Fig. 2) does not influence changes in the assessed constructs.

3.2. Changes at the measures level

The standardized changes computed for the complete set of measures are represented in Fig. 5 (left panel for the intelligence tests and right panel for the cognitive tasks).

There were no significant differences between the training and control groups on the measures of fluid intelligence, although for the RAPM it was at a trend level ($p = .06$). The training group showed a greater change for the DAT-AR, although the difference between groups was not significant. There was a large change for both groups in the highly speeded Gf test (PMA-R) amounting to 1 SD.

With respect to crystallized intelligence measures, the changes for the training and control groups were all small. In

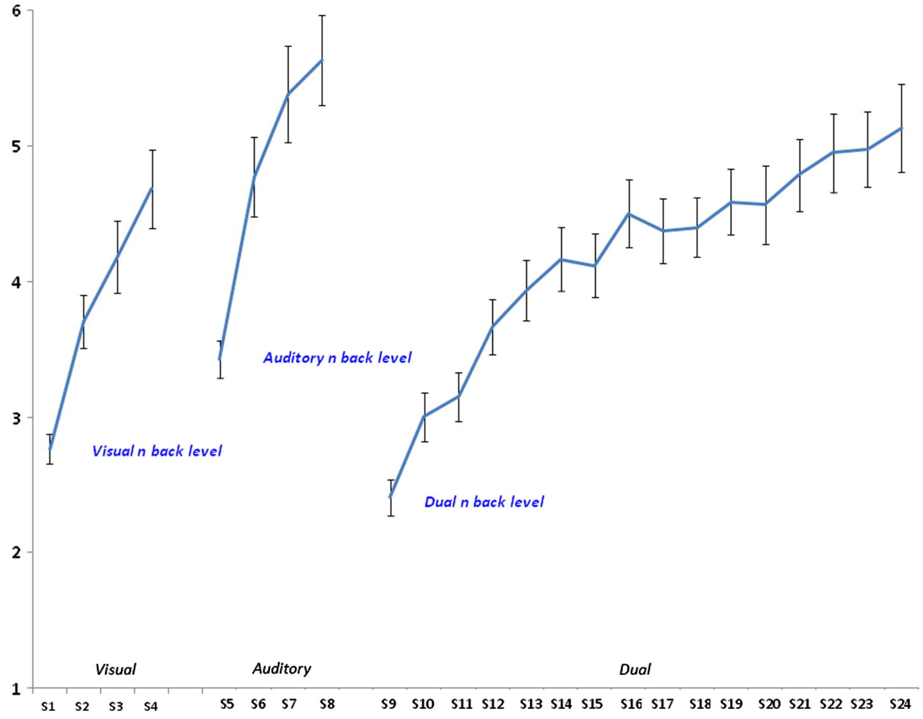

Fig. 2. Average n-back level achieved (Y-axis) by the training group (N = 28) across the visual, auditory, and dual sessions. S = session.

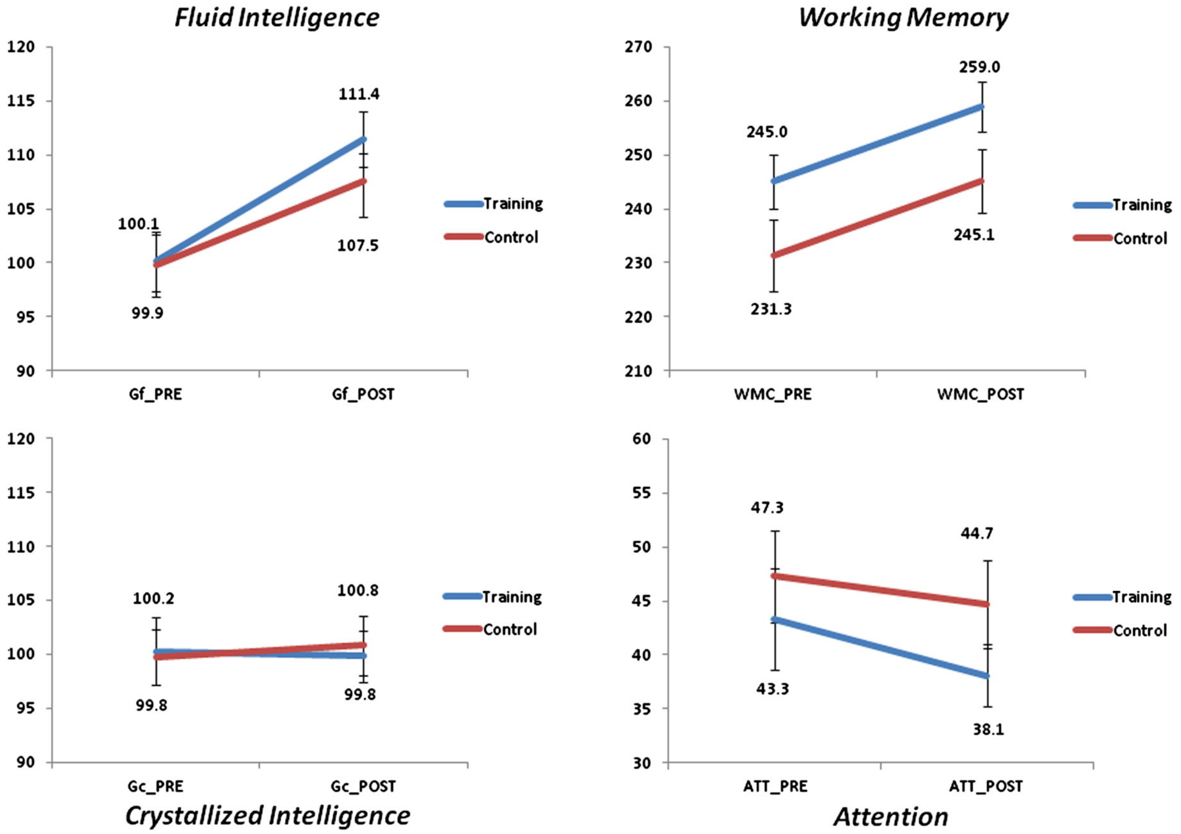

Fig. 3. Scores in the pretest and posttest sessions for the training (N = 28) and control (N = 28) groups at the construct level. The values are recovered from Table A.1. Values for fluid intelligence and crystallized intelligence are derived from the IRT transformations (Appendix A.3).

addition, the difference between groups on the observed standardized changes was not statistically significant.

The findings for working memory measures were interesting. Firstly, there was a significant difference for the reading span task between the training and control groups ($p = .01$) favoring the former. Secondly, the difference for the dot matrix task was also significant ($p = .04$) and it again favored the training group. Thirdly, the difference in the computation span task was not significant. Therefore, two out of three working memory measures showed statistically significant changes favorable to the training group.

Finally, two out of three attention control measures revealed non-significant results. The changes were generally small for the training and control groups in the three attention control measures. However, the difference for the spatial Simon task was significant and it favored the training group ($p = .01$).

4. Discussion

The main finding is that the large improvements in the challenging adaptive cognitive training program based on the n-back task (Fig. 2) do not evoke greater changes than those observed for a passive control group in fluid-abstract intelligence and crystallized intelligence, or in working memory capacity and attention control at the construct level. This happens even when average n-back performance across the training sessions shows significant correlations with crystallized intelligence, working memory capacity, and especially, fluid intelligence. This result conflicts with previous reports supporting a positive effect of this short-term adaptive cognitive training over fluid intelligence performance (Jaeggi et al., 2008, 2010, 2011, in press; Stephenson & Halpern, 2013).

At the construct level, the findings reported here seem to be consistent with Chooi and Thompson's (2012) and Redick et al.'s (2012). Chooi and Thompson (2012) assessed intelligence changes from the pretest to the posttest after training on the adaptive n-back dual task measuring the constructs of verbal intelligence, perceptual intelligence, and mental rotation, taking the VPR model as a frame of reference (Johnson & Bouchard, 2005). However, this study found decreased scores at the posttest for verbal intelligence and perceptual intelligence in the three tested groups (training, active control, and passive control). For mental rotation, there were increments for both control groups and hardly any change for the training group. No changes were found for the RAPM test (Gf), the Mill Hill test (Gc), and the operation span task (working memory) across the three groups. Note that scores for the active and passive control groups were almost identical for the assessed psychological constructs, which contradicts Shipstead et al.'s (2012) reservations with respect to the use of passive control groups within this research context (see further discussions).

Redick et al. (2012) achieved similar conclusions. As discussed above, this study assessed fluid intelligence, crystallized intelligence, working memory, multitasking, and processing speed. They failed to find any difference at the constructs or at the measures levels among their trained, active, and passive control groups. Further, contrary to what was found here, Redick et al. (2012) report a lack of significant correlations between average n-back performance and the measured psychological constructs. Together with the low average n-back performance achieved by their trained participants, reservations can be raised regarding the straight comparison of their findings and those observed in the present study.

We suggest that the studies by Chooi and Thompson (2012) and Redick et al. (2012) may suffer measurement problems. It

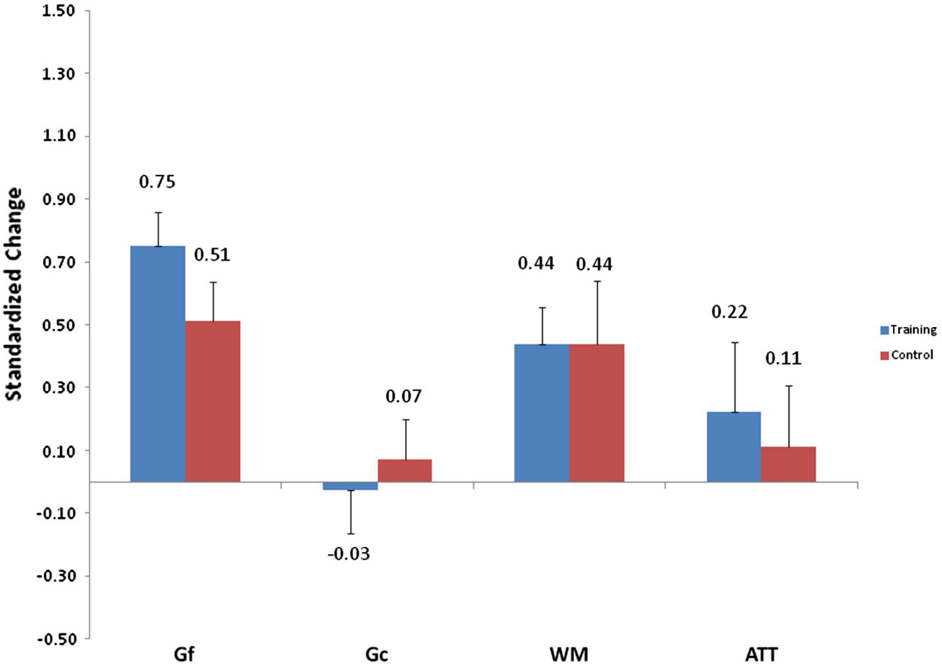

Fig. 4. Standardized change [posttest − pretest / SD at the pretest] of the training and control groups in the assessed psychological constructs (Gf = fluid intelligence, Gc = crystallized intelligence, WMC = working memory capacity, ATT = attention control). ANCOVA results: [Gf] $F(1,53) = 2.380$; $p = .06$; $\eta^2 = .043$, [Gc] $F(1,53) = .267$; $p = .30$; $\eta^2 = .005$, [WMC] $F(1,53) = 1.088$; $p = .15$; $\eta^2 = .020$, [ATT] $F(1,53) = 1.429$; $p = .12$; $\eta^2 = .026$.

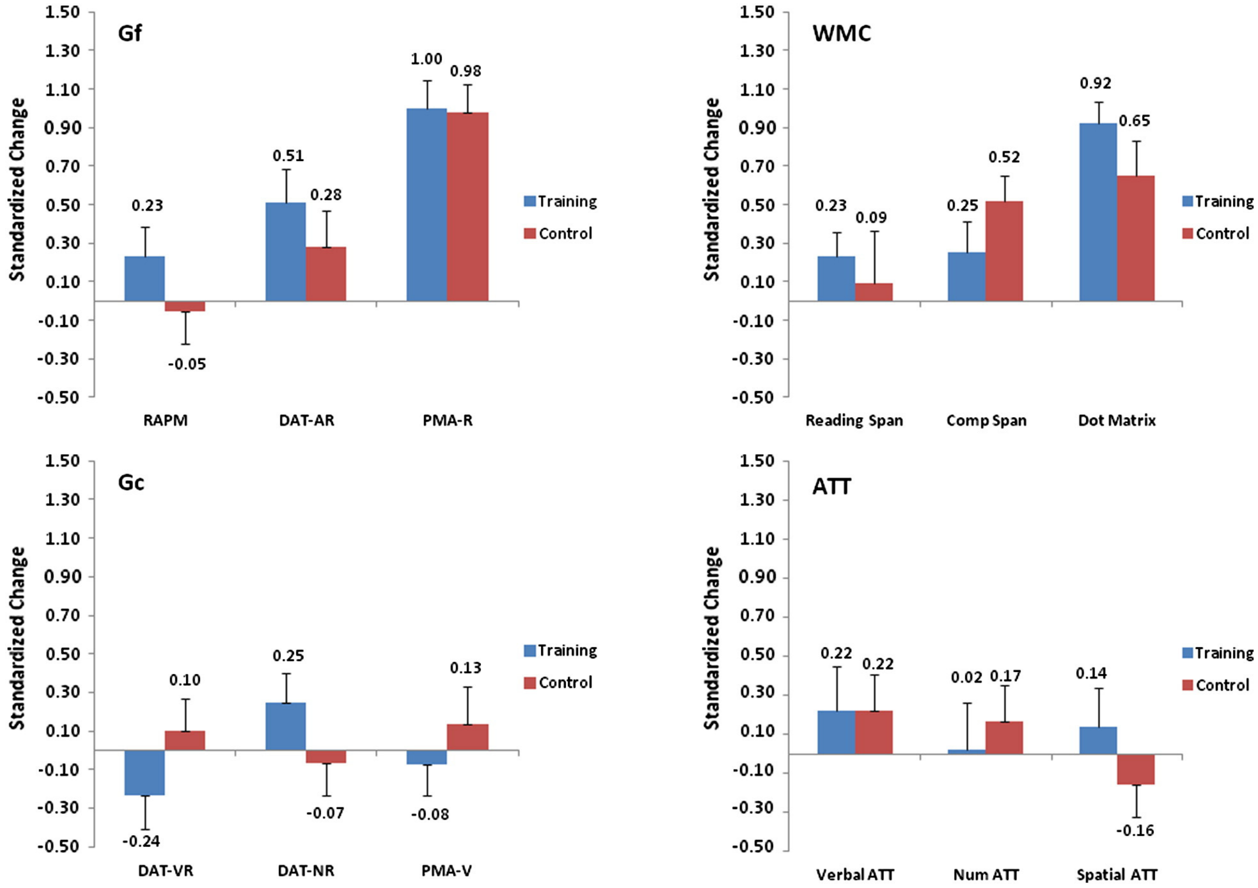

Fig. 5. Standardized change [posttest − pretest / SD at the pretest] of the training and control groups in the administered measures organized by tapped construct: Gf = fluid intelligence (RAPM = Raven Advanced Progressive Matrices Test, DAT-AR = abstract reasoning, PMA-R = inductive reasoning); Gc = crystallized intelligence (DAT-VR = verbal reasoning, DAT-NR = numerical reasoning, PMA-V = vocabulary); WMC = working memory capacity; ATT = attention control (verbal ATT = vowel–consonant, numerical ATT = odd–even, spatial ATT = right–left). ANCOVA results: [RAPM] $F(1,53) = 2.340$; $p = .06$; $\eta^2 = .042$, [DAT-AR] $F(1,53) = .468$; $p = .25$; $\eta^2 = .009$, [PMA-R] $F(1,53) = .304$; $p = .29$; $\eta^2 = .006$, [DAT-VR] $F(1,53) = 1.162$; $p = .14$; $\eta^2 = .021$, [DAT-NR] $F(1,53) = .538$; $p = .23$; $\eta^2 = .010$, [PMA-V] $F(1,53) = .133$; $p = .36$; $\eta^2 = .002$, [Reading span] $F(1,53) = 5.067$; $p = .01$; $\eta^2 = .087$, [Computation span] $F(1,53) = .976$; $p = .16$; $\eta^2 = .018$, [Dot matrix] $F(1,53) = 2.854$; $p = .04$; $\eta^2 = .051$, [V-ATT] $F(1,53) = .039$; $p = .44$; $\eta^2 = .001$, [N-ATT] $F(1,53) = .019$; $p = .44$; $\eta^2 = .000$, [S-ATT] $F(1,53) = 6.257$; $p = .007$; $\eta^2 = .106$.

is really surprising to see no changes at all across three measurement time points (Redick et al., 2012) or even worse performance after training (Chooi & Thompson, 2012; Redick et al., 2012). As noted by Jaeggi et al. (in press), splitting standard tests in half or thirds might reduce reliability and validity which, in turn, may lead to a loose in sensitivity.

Re-test or practice effects are well-known and largely documented. Back in 1930s, Anastasi (1934) published a seminal study showing gains ranging from $d = 0.2$ to $d = 1.1$ by the mere effect of practice. Reeve and Lam (2005) reported practice effects ranging from $d = 0$ to $d = .85$. Colom et al. (2010) found changes ranging from $d = 0$ to $d = 0.73$. This is what Jensen (1998) concludes with respect to these re-test or practice effects: "when the same test, or an equivalent or parallel form of the test, is administered to persons on two separate occasions, there is usually an increase in scores, called a 'practice effect'. Typically, the initial gain amounts to about three to six points on the IQ scale" (pages 314–315). Failing to find this expected practice effect seems odd. Nevertheless, perhaps the set of items administered in their re-test sessions were more difficult than those administered in the pretest sessions. This is suggested by Chooi and Thompson (page 537) with respect to their verbal fluency and perceptual speed tests, but it is unlikely applicable to the Redick et al. (2012) study because participants were re-tested on two occasions.

A close look at the findings reported in the present study reveals several noteworthy issues. First, the standardized change in fluid intelligence for the training group is almost twice as big as that observed for the control group (Figs. 3 and 4). As reported, the computed analysis of covariance approached the fixed level for statistical significance. The inspection of the specific measures of fluid intelligence is also revealing (Fig. 5). The screening versions of the RAPM test were administered with a very liberal time limit, so the reservation raised by Moody (2009) in this regard seems inappropriate for the present study. The change for the training group was greater than for the control group (again significant at a trend level). The training group also showed a greater change in the screening version of the abstract reasoning (AR) subtest from the DAT battery (also administered with a liberal time limit). For the highly speeded inductive reasoning (R) subtest from the PMA, both groups showed a large (and identical) change from the pretest to the posttest.

Second, Shipstead et al. (2012) raised reasonable doubts regarding the use of passive control groups in these cognitive training studies. However, the results reported here with respect to crystallized intelligence suggest that the type of factors enumerated by these researchers (Hawthorne effect, etc.) were not operative in the present study. The standardized changes for the control group in the Gc construct and in their specific measures are parallel to those that were observed for the training group. This is also consistent with the results reported by Chooi and Thompson (2012) and Redick et al. (2012) as discussed above. Note finally that these types of factors are much less relevant than generally assumed (Adair, Sharpe, & Huynh, 1989; Kompier, 2006; Wickstrom & Bendix, 2000).

Third, Jaeggi et al. (2008) failed to find changes in a working memory measure (reading span) after the application of the cognitive training on the adaptive dual n-back (although, interestingly, they found significant changes for digit span, a pure short-term memory measure). The same lack of change was noted by Chooi and Thompson (2012) and by Jaeggi et al. (2010) for the operation span task, as well as by Redick et al. (2012) for the symmetry and running span tasks. Here, we have shown that at the construct level, the standardized improvement is almost the same for the training and control groups. However, two out of three working memory measures showed statistically significant differences between groups. Dot matrix and reading span improvements were substantially higher for the training group than for the control group. This result is reversed for the computation span task. Note that this task parallels the operation span task and our results are consistent with those found by Jaeggi et al. (2010) for their passive control group. Averaging the two types of results in the present study produces a null difference at the construct level for working memory capacity, which reinforces the caution note regarding construct heterogeneity.

Finally, for attention control the general findings were similar to those found for crystallized intelligence: there was a very small standardized change for both groups at the construct level and this also was observed for the specific attention measures. The exception was for spatial attention, for which the training group showed a standardized change significantly different to the change observed for the control group.

Taken together, the results for the twelve measures administered at the pretest and posttest sessions suggest that the cognitive intervention used here may enhance visuospatial processing (also consistent with Jaeggi et al., in press). The visuospatial fluid measures (RAPM and abstract reasoning — DAT-AR), along with spatial working memory (dot matrix), and spatial attention control (Simon task) showed the greatest difference between the training and control groups (Fig. 5) favoring the former. This observation is reinforced by the negative results for the crystallized-verbal measures, computation span (working memory), and the verbal and numerical attention control tasks. The reading span task seems like an exception to this general pattern. However, it should be noted that there is a clear spatial requirement for this working memory task (see Appendix A.2): participants must recall the displayed letters (secondary task) according to their 'position' in the alphabet ignoring their serial order in the sequence. Further, the auditory n-back condition was based on the updating of the set of letters and this might have some positive specific impact here.

The meta-analysis reported by Melby-Lervåg and Hulme (2012) supports the positive result for these visuospatial processing skills. These researchers analyzed twenty-three studies finding reliable short-term and specific increments in working memory skills after cognitive training. Note also that the meta-analyses published by Hindin and Zelinski (2012) and Uttal et al. (2013) found small/medium positive effect sizes for cognitive training in terms of improvement in non-trained domains. Results reported by Rudebeck, Bor, Ormond, O'Reilly, and Lee (2012) and von Bastian and Oberauer (2013) are also consistent with the findings reported here.

The study by Stephenson and Halpern (2013) deserves a special comment. As noted at the Introduction section, Jaeggi et al.'s (2008, 2010) main findings were replicated by these researchers. Nevertheless, significant improvements in fluid intelligence tests were observed after the visuospatial short-term memory (STM) training program (in addition to those observed for the dual n-back training program). This led to the conclusion that "STM training had an effect because the STM training enhanced the shared short-term storage component that influences Gf. The constructs STM, WMC, executive functioning, attention, and Gf do have a common factor: short-term storage capacity. Perhaps, what the cognitive training is truly doing is expanding participants' limited capacity that all of the constructs have in common" (page 354). This nicely fits the main theoretical background framing the present study, namely, both fluid intelligence and working memory capacity require the reliable preservation of the relevant information in the short-term (Colom et al., 2006, 2008; Martínez et al., 2011). If adaptive working memory/short-term memory training promotes skills relevant for the reliable temporary storage of relevant information, then fluid intelligence and working memory capacity scores will be higher for a trained than for a control group. Indeed, (a) the recent report by Jaeggi et al. (in press) suggests that executive updating processes may support the relationships among these constructs and (b) Martínez et al. (2011) demonstrated a near-perfect correlation, at the latent variable level, among short-term memory, executive updating, working memory, and fluid intelligence.

In closing, the main conclusion is that the short-term challenging adaptive cognitive training based on the n-back task does not increase performance in fluid intelligence at the construct level. Nevertheless, post-hoc analyses done at the measures level suggest further research to determine if the administered cognitive training may enhance visuospatial processing skills.

Acknowledgments

This research was supported by Grant PSI2010-20364 (Ministerio de Ciencia e Innovación, Spain). FJR is also supported by BES-2011-043527 (Ministerio de Ciencia e Innovación, Spain). KM is also supported by AP2008-00433 (Ministerio de Educación, Spain). MB was funded by grant "Alianza 4 Universidades" program (A4U-4-2011).

Appendix A

A1.

Table A.1
Descriptive statistics for the demographic variables and performance on the cognitive measures for the two groups of participants (training and control). IRT = Item Response Theory scores. SD = standard deviation.

	Training group (N = 28)				Control group (N = 28)			
	Mean	IRT-mean	SD	IRT-SD	Mean	IRT-mean	SD	IRT-SD
Age	18.04		0.9		18.2		1.2	
General intelligence	101.00		16.4		101.00		16.00	
Pretest								
Gf	31.89	100.13	7.09	14.37	31.93	99.87	7.67	15.87
RAPM	11.82	100.88	2.65	14.04	11.54	99.12	2.97	16.11
DAT-AR	11.32	98.77	3.53	14.67	11.86	101.23	3.62	15.49
PMA-R	8.75	100.65	3.00	15.61	8.54	99.35	2.78	14.63
Gc	38.25	100.24	8.81	16.54	37.71	99.76	7.55	13.59
DAT-VR	13.14	101.85	3.42	16.26	12.39	98.15	3.01	13.67
DAT-NR	7.32	96.68	3.01	15.40	8.61	103.32	2.78	14.09
PMA-V	17.79	102.04	4.13	15.33	16.71	97.96	4.02	14.65
Working memory	245.04		26.67		231.32		35.12	
Reading span	128.50		14.58		119.36		21.02	
Computation span	63.50		14.54		60.04		17.08	
Dot matrix	53.04		4.60		51.93		7.42	
Attention	43.33		24.65		47.29		22.54	
Verbal	59.00		36.80		57.71		40.39	
Numerical	37.39		37.92		47.54		36.26	
Spatial	33.61		48.37		36.61		38.45	
Posttest								
Gf	37.25	111.45	6.23	13.45	35.46	107.55	8.26	17.48
RAPM	11.79	104.34	2.27	12.27	10.64	98.33	3.25	17.82
DAT_AR	13.64	106.39	3.30	14.27	13.36	105.42	4.00	17.05
PMA_R	11.82	116.25	2.21	13.20	11.46	114.03	2.32	13.37
Gc	35.68	99.84	6.22	12.51	36.00	100.81	7.42	14.74
DAT-VR	12.32	98.31	2.74	12.87	12.57	99.69	3.47	16.28
DAT-NR	8.82	100.39	2.68	13.79	9.18	102.28	3.07	15.72
PMA-V	14.54	100.91	2.76	12.27	14.25	99.94	3.87	16.92
Working memory	258.93		24.59		245.14		31.60	
Reading span	132.82		10.51		121.07		19.96	
Computation span	67.46		14.04		68.21		14.54	
Dot matrix	58.64		5.14		55.86		6.92	
Attention	38.08		15.22		44.68		21.31	
Verbal	50.43		22.87		49.21		19.27	
Numerical	36.50		35.91		41.29		36.41	
Spatial	27.32		20.79		43.54		28.26	

A.2

Table A.2

Detailed description of intelligence tests and cognitive tasks administered in the present study.

Tests/tasks	Description
Fluid-abstract intelligence (Gf) evaluates the achieved complexity level in problems at which previous knowledge is useless.	
RAPM	The RAPM comprises a matrix figure with three rows and three columns. Among eight possible alternatives the one completing the 3 × 3 matrix figure must be chosen. The screening version comprising odd items only was administered in the pretest, whereas the even items were administered in the posttest.
DAT-AR	DAT-AR is a series test based on abstract figures. Successive figures follow a given rule, so the one continuing the series must be chosen from several alternatives. The screening version comprising odd items only was administered in the pretest, whereas the even items were administered in the posttest.
PMA-R	PMA-R comprises letter series items. The rule (or rules) underlying a given sequence must be extracted for selecting the correct alternative. The screening version comprising odd items only was administered in the pretest, whereas the even items were administered in the posttest.
Crystallized-verbal intelligence (Gc) is supported by ability to solve academic subjects such as reading and math.	
DAT-VR	DAT-VR is based on sentences stated like an analogy. The first and last words from the sentence are missing, and a pair of words completing the sentence must be selected. The screening version comprising odd items only was administered in the pretest, whereas the even items were administered in the posttest.
DAT-NR	DAT-NR consists of quantitative reasoning problems. The screening version comprising odd items only was administered in the pretest, whereas the even items were administered in the posttest.
PMA-V	PMA-V is a synonym test based on the meaning of words that must be evaluated against a given model word. The screening version comprising odd items only was administered in the pretest, whereas the even items were administered in the posttest.
Working memory capacity (WMC) captures the ability for temporarily store varied amounts of information while solving a concurrent processing requirement (the score for the WMC tasks is the number of hits in the verification and recalling tasks).	
Reading span	In the reading span task, participants verify if a set of sentences sequentially displayed make or make no sense. Each display includes a sentence and a to-be remembered capital letter. Sentences are 10–15 words long. At the end of a given set, participants recall, according to their position in the alphabet and irrespective of their serial order, each letter from the set. Set sizes range from 3 to 7 sentence/letter pairs per trial, for a total of 12 trials (5 levels × 3 trials = 15 trials total). Difficulty levels were randomly presented.
Computation span	The computation span task includes a verification task and a recall task. 6 s are allowed to see the math equation without a time limit for verifying its accuracy. The displayed solution, irrespective of its accuracy, must be serially remembered at the end of a given set. Each math equation includes two operations using digits from 1 to 10. The solutions are always single-digit numbers. Trials range from three to seven equations/solutions (5 levels × 3 trials each = 15 trials total). Difficulty levels were randomly presented.
Dot matrix	In the dot matrix task, a matrix equation must be verified and a dot location displayed in a five × five grid must be retained. The matrix equation is presented during a maximum of 4.5 s for adding or subtracting simple line drawings. Once the response is given, the grid comprising the to-be remembered dot is displayed for 1.5 s. After a given set of equation–grid pairs, the grid spaces that contained dots must be recalled clicking with the mouse on an empty grid. Trials increase in size from three to five equations and dots (3 levels × 3 trials = 9 trials total). Difficulty levels were randomly presented.
Attention is a cognitive function for focusing available mental resources and here we consider the control of automatic responses (inhibition). [The compatibility effect (reaction time for the incompatible trials minus reaction time for the compatible trials) was the dependent measure].	
Attention control	Attention control is measured here by means of verbal and quantitative versions of the flanker task and a version of the Simon task. The verbal and quantitative tasks require deciding, as fast as possible, if the letter/digit presented in the center of a set of three letters/digits is vowel/odd or consonant/even. The target (e.g. vowel/odd) can be surrounded by compatible (e.g. vowel/odd) or incompatible (e.g. consonant/even) letters/digits. The spatial task requires deciding if an arrow (horizontally depicted) points to the left or to the right 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. There are a total of 32 practice trials and 80 experimental trials. Half of the trials are compatible and they are randomly presented across the entire session.

Note: Four out of six intelligence tests were applied without severe time constraints. For the RAPM, there was more than 1 min per item (20 min for 18 items). For DAT-AR, DAT-NR and DAT-VR, there were approximately 30 s per item (10 min for 20 items). For the speeded tests (PMA-R and PMA-V), there were between 5 and 12 s per item (PMA-R: 3 min for 15 items and PMA-V: 2 min for 25 items).

A.3. Application of item response theory (IRT) for equating the difficulty levels of pretest and posttest measures of intelligence

A.3.1. Rationale

Analyzing raw or composite scores, we observed that for some tests (especially crystallized measures) performance decreased in the posttest. We reasoned that this unexpected trend may result from differences in the difficulty level of the administered items in the pretest (odd items) and posttest (even items) sessions. Further, we thought that some positive changes may be attributed to these differences in difficulty. We checked and confirmed this possibility analyzing several comparable samples that completed the full versions of the tests administered in the present study. Results are shown in Table A.3.1.

Mean differences between the odd and even items were significant ($p < 0.001$ for all the tests, excluding the DAT-VR) which implies that pretest (odd) and posttest (even) scores must not be directly compared. For fixing this problem we applied an item response theory (IRT) scoring procedure. In item response theory, the ability (called θ) is estimated as the trait that maximizes the likelihood of the response pattern. As a result, IRT models may produce pretest and posttest θ scores that are more independent of the particular set of administered items (invariance property; Hambleton & Swaminathan, 1985).

Table A.3.1
Mean (SD) and statistical test for differences between scores in odd and even items.

	N	Odd	Even	t(gl); *p*	α Coefficient		
					Odd	Even	Total
Gc							
DAT-VR	416	13.25 (3.29)	13.12 (3.05)	t(415) = 1.03; *p* = .302	.703	.634	.802
DAT-NR	195	9.03 (3.40)	9.77 (3.49)	t(194) = −4.95; *p* < .001	.704	.755	.852
PMA-V	325	16.35 (3.47)	13.49 (2.75)	t(324) = 22.91; *p* < .001	.782	.655	.847
Gf							
RAPM	327	11.98 (2.83)	11.28 (2.80)	t(326) = 5.49; *p* < .001	.654	.660	.795
DAT-AR	327	12.87 (2.80)	13.43 (3.70)	t(326) = −4.11; *p* < .001	.765	.775	.871
PMA-R	327	9.34 (2.63)	9.64 (2.64)	t(326) = −3.94; *p* < .001	.725	.776	.868

One typical example of IRT scoring is a computerized adaptive test, in which although each examinee receive a different set of items (fitted in difficulty to their observed performance) θ scores are estimated in the same metric.

Here, we applied item response theory to parcel scores instead of doing so for the specific items. Analyses revealed that the factorial structure for the items was bidimensional due to time constraints (the last item in the test loaded in a second factor that may be interpreted as a "speed factor"). Application of item response theory requires unidimensionality. However, Reckase, Ackerman, and Carlson (1988) have shown that sets of items that measure the same composite of abilities may meet the unidimensionality assumption. In our case, item parcels were constructed to measure the same composite of power and speed abilities. Then, parcels were treated as unidimensional polytomous items in the analysis. One additional advantage of item parceling is that the number of variables is reduced.

A.3.2. Method

A.3.2.1. Calibration samples. For applying IRT scoring to the data obtained in the present study, four independent samples were analyzed for item parameter calibration (N = 416, for calibrating the DAT-VR; N = 195 for DAT-NR; N = 327 for RAPM, DAT-AR, and PMA-R; N = 325 for PMA-V). The analyzed samples were strictly comparable (university undergraduates).

A.3.2.2. Item parceling. As noted above, specific items were grouped in parcels. Odd and even items were parceled separately. These parcels balance power and speed in the same way. For example, for the DAT-NR test, the 20 odd items were sequentially assigned to the four five-item parcels (1st odd item to the first facet, 2nd odd item to the second facet, 3th odd item to the third facet, 4th odd item to the fourth facet, 5th odd item to the first facet, and so on). Thus, resulting parcels are "unidimensional" (although the measured factor will be a balanced composite of power and speed). The numbers of five-item parcels were: 10 (PMA-V), 8 (DAT-VR, DAT-NR and DAT-AR), and 6 (PMA-R). For the RAPM, 12 three-item parcels were constructed.

A.3.2.3. Unidimensionality and local independence. Before applying the IRT model, unidimensionality and local independence assumptions were tested. Fit of the unidimensional models was assessed using the root mean square error of approximation (RMSEA) and the comparative fit index (CFI). Values close to .95 for CFI and below .06 for RMSEA indicate a good fit (Hu & Bentler, 1999). Polychoric correlations were analyzed using weighted least squares with adjustments for the mean and variance (WLSMV) estimator in MPLUS 7 (Muthén & Muthén, 2012). We further examined the percentage of variance explained by the first factor (at least 20% is desirable) and item loadings (loadings larger than .20 are desirable). The Scree test was inspected as a complementary tool. For assessing local dependence between items, the residual correlation matrix was inspected. High residuals (e.g., larger than 0.2) or high Modification Indices may indicate a local dependence problem (Reeve, Hays, Bjorner, et al., 2009).

A.3.2.4. Item response theory calibration. The polytomous graded response model (Samejima, 1969) was fitted to the data with the IRTPRO 2.1 program (Cai, du Toit, & Thissen, 2011). Applying this model assumes that the probability of scoring *k* or larger on a parcel, *Xj*, is an increasing function of θ and follows the logistic model:

$$P(X_j \geq k|\theta) = \frac{1}{1+\exp\left(-a_j\left(\theta-b_{jk}\right)\right)}$$

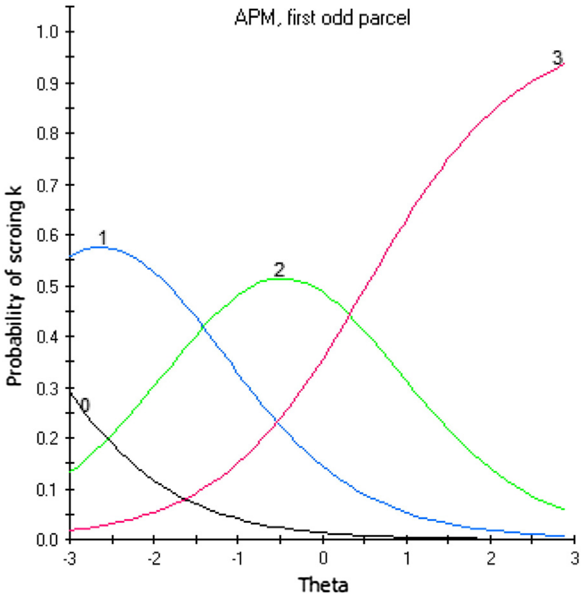

Fig. A.3.1. Probability of scoring *k* (*k*: 0, 1, 2 and 3) for the RAPM first odd parcel (sum of correct responses to the items 1, 13, and 25).

Table A.3.2
Summed scores to θ scores conversion tables for odd and even parts of the RAPM.

Odd part conversion table		Even part conversion table	
Summed score	θ	Summed score	θ
0	−3.528	0	−3.395
1	−3.191	1	−3.040
2	−2.860	2	−2.696
3	−2.539	3	−2.370
4	−2.230	4	−2.062
5	−1.934	5	−1.769
6	−1.648	6	−1.485
7	−1.370	7	−1.207
8	−1.097	8	−0.930
9	−0.826	9	−0.654
10	−0.556	10	−0.378
11	−0.286	11	−0.101
12	−0.013	12	0.183
13	0.265	13	0.477
14	0.553	14	0.787
15	0.858	15	1.119
16	1.189	16	1.479
17	1.555	17	1.880
18	1.967	18	2.280

where a_j is the discrimination parameter and b_{jk} are the extremity parameters that depends on k ($b_{j1} \leq b_{j2} \ldots \leq b_{jK-1}$, being K the maximum score). The probability of scoring k is obtained as a difference:

$$P(X_j = k|\theta) = P(X_j \geq k|\theta) - P(X_j \geq k+1|\theta).$$

Fig. A.3.1 plots the probability of scoring k as function of θ for the first parcel of the RAPM. The lower the ability, the larger is the probability of scoring higher in the parcel. In IRTPRO, maximum marginal likelihood estimation with an EM algorithm (Bock and Aitkin, 1981) was used to estimate IRT item and person parameters for unidimensional models.

Goodness of fit of item response models was checked using the computer macro, IRTFIT (Bjorner, Smith, Stone, & Sun, 2007). We compute G^{*2} statistics for each item. These statistics compare expected and observed frequencies of item category responses for various levels of θ and quantify the differences between expected and observed responses. Significance levels are obtained by a Monte Carlo re-sampling procedure (Stone & Zhang, 2003).

A.3.2.5. Summed-score expected a posteriori estimates. Summed score to θ scores conversion tables were obtained separately for odd and even parts of each test, using obtained item parameter estimates. These tables were used to obtain *summed-score expected a posteriori estimates* (Cai, Du Toit, & Thissen, 2011; p. 160). One SSEAP estimate score is the expected θ for one obtained summed-score S [$E(\theta|S)$]. One advantage of these IRT estimates is that they do not require to know the pattern of item responses since we only need the summed score to obtain θ (Thissen & Wainer, 2001).

The conversion table was used to obtain θ estimates for participants of the present study. One example of conversion table for the RAPM is shown in Table A.3.2.

Table A.3.2 shows that the same summed score implies a larger SSEAP θ score (because the even part of the RAPM is more difficult than the odd part).

A.3.3. Results

A.3.3.1. IRT assumptions: unidimensionality and local independence. Goodness of fit indexes for the unidimensional models are shown in Table A.3.3. Unidimensionality and local independence was supported for all tests. CFI values were larger than 0.95 (between 0.965 and 0.997). Likewise, RMSEA values are lower than 0.06 for four scales (DAT-VR, DAT-NR, RAPM, and DAT-AR) and lower than 0.08 for the remaining (PMA-V and PMA-R), suggesting a reasonable fit for the unidimensional model. Furthermore, the Scree test supported the one-factor solution and percentages of variance accounted for by the first factor varied between 26% and 44% depending on the scale. Finally, examination of the residual correlations indicated very minor local dependence (residuals were not greater than 0.15). For DAT-VR, DAT-AR, PMA-R, and RAPM, only 5% of the residuals were greater than .10.

A.3.3.2. IRT calibration and fit. No items were found to misfit the GRM ($p > 0.01$). Range and average probability values for the G^{*2} statistics are shown in Table A.3.3. Fig. A.3.2 shows the summed score to θ scores conversion figures for odd and even parts of each test. As can be seen, the highest correction in the IRT scoring is carried out for the PMA-V test, but there are non-negligible corrections for DAT-NR and RAPM. For PMA-V and RAPM, correct responses are more positively weighted in the (more difficult) part (even items), whereas for the DAT-NR the correct responses are more positively weighted in the odd items.

Table A.3.3
Goodness of fit for unidimensional models.

	One factor						GRM	
	RMSEA	CFI	% variance	Lowest loading	Residual >.1	Largest residual	Range P (G^{*2})	Average p (G^{*2})
Gc								
DAT-VR	.035	.992	26	.411	4%	−.102	.44–.91	.66
DAT-NR	.058	.992	34	.377	7%	−.138	.30–.97	.74
PMA-V	.067	.987	44	.516	7%	−.136	.07–.88	.38
Gf								
RAPM	.046	.965	30	.451	6%	−.145	.21–.94	.59
DAT-AR	.034	.997	36	.674	0%	.087	.15–.93	.49
PMA-R	.073	.996	35	.789	0%	−.057	.04–.82	.45

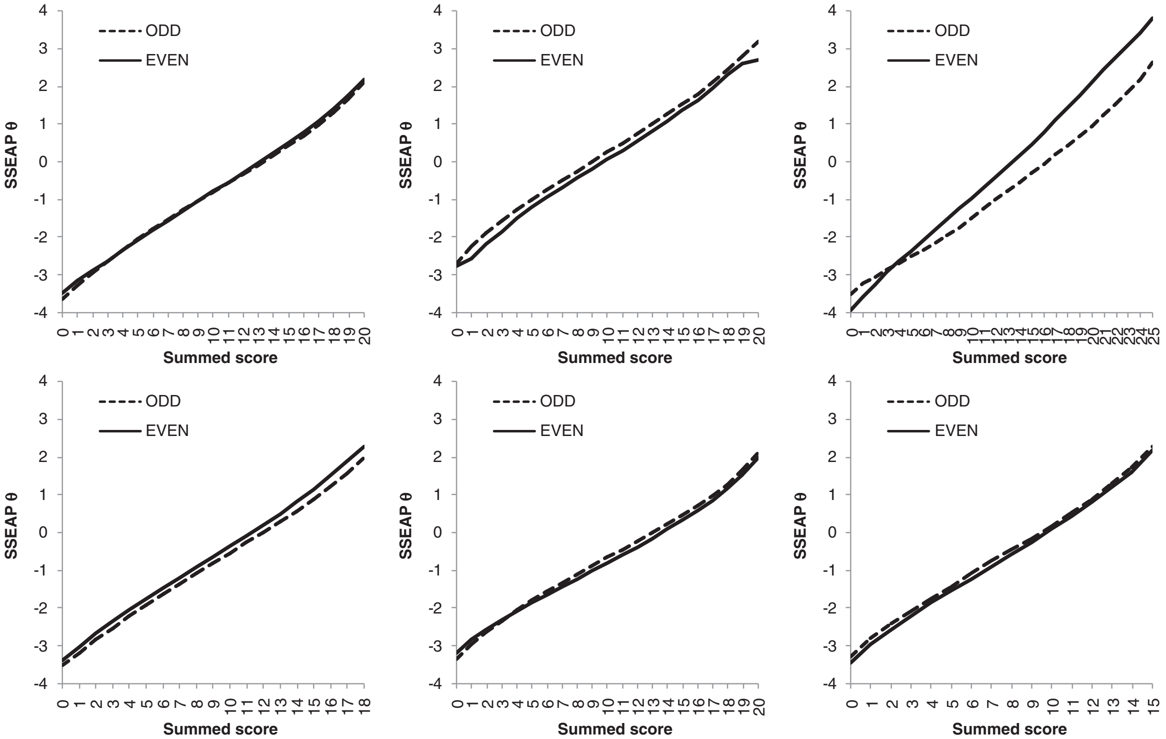

Fig. A.3.2. Summed score to θ scores conversion figures for odd and even parts of each test. In the top panel, from left to right (DAT-VR, DAT-NR and PMA-V). In the bottom panel, from left to right (RAPM, DAT-AR and PMA-R).

A.4

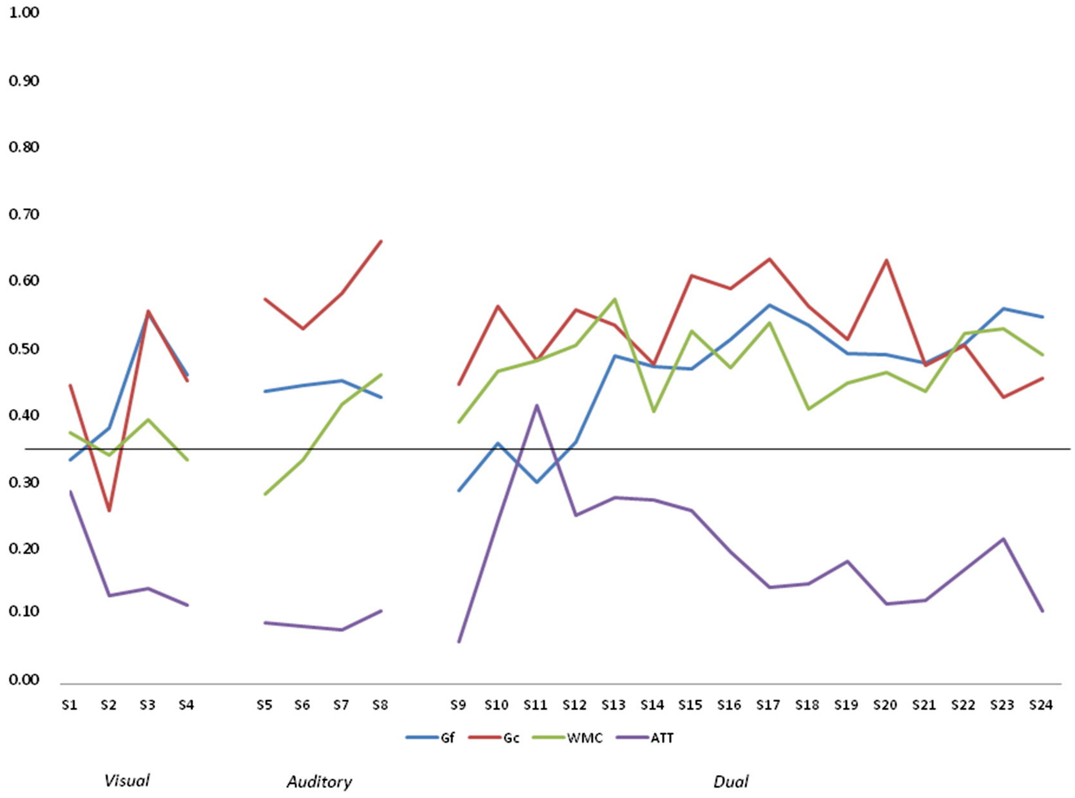

Fig. A.4. Correlation between pretest intelligence/cognitive factors and achieved n-back levels across the full range of training sessions.

References

Adair, J. G., Sharpe, D., & Huynh, C. L. (1989). Hawthorne control procedures in educational experiments: A reconsideration of their use and effectiveness. *Review of Educational Research*, *59*(2), 215–228.

Anastasi, A. (1934). Practice and variability. *Psychological Monographs*, *45*, 5.

Bjorner, J. B., Smith, K. J., Stone, C., & Sun, X. (2007). *IRTFIT: A macro for item fit and local dependence tests under IRT models.* Lincoln: Quality Metric, Inc.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [computer software and manual].* Chicago, IL: Scientific Software International.

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.

Cattell, R. B. (1987). *Intelligence: Their structure, growth and action.* Amsterdam: North-Holland.

Chooi, W., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*, 531–542.

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

Colom, R., Quiroga, Mª. A., Shih, P. C., Martínez, K., Burgaleta, M., Martínez-Molina, A., et al. (2010). Improvement in working memory is not related to increased intelligence scores. *Intelligence*, *38*, 497–505.

Colom, R., Quiroga, Mª.Á., Solana, A. B., Burgaleta, M., Román, F. J., Privado, J., et al. (2012). Structural changes after videogame practice related to a brain network associated with intelligence. *Intelligence*, *40*, 479–489.

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A re-analysis of key studies. *Memory & Cognition*, *34*, 158–171.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory.* : Psychology Press.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Gottfredson, L., et al. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*, 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.

Hambleton, R. K. y, & Swaminathan, H. (1985). *Item response theory. Principles and applications.* Boston, MA: Kluwer Nijhoff Publishing.

Hindin, S. B., & Zelinski, E. M. (2012). Extended practice and aerobic exercise interventions benefit untrained cognitive outcomes in older adults: A meta-analysis. *Journal of the American Geriatrics Society*, *60*(1), 136–141. http://dx.doi.org/10.1111/j.1532-5415.2011.03761.x.

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

Hunt, E. B. (1995). *Will we be smart enough?* A cognitive analysis of the coming workforce. New York: Russell Sage Foundation.

Hunt, E. B. (2011). *Human intelligence.* Cambridge: Cambridge University Press.

Irwing, P., Hamza, A., Khaleefa, O., & Lynn, R. (2008). Effects of abacus training on the intelligence of Sudanese children. *Personality and Individual Differences*, *45*(7), 694–696.

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

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *PNAS*, *108*, 10081–10086.

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2013). The role of individual differences in cognitive training and transfer. *Memory & Cognition* (in press).

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning — Implications for training and transfer. *Intelligence*, *38*, 625–635.

Jaušovec, N., & Jaušovec, K. (2012). Working memory training: Improving intelligence — Changing brain activity. *Brain and Cognition*, *79*, 96–106.

Jensen, A. R. (1998). *The g factor.* New York: Praeger.

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

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of Educational Research*, *78*, 1142. http://dx.doi.org/10.3102/0034654308327416.

Kompier, M. A. (2006). The "Hawthorne effect" is a myth, but what keeps the story going? *Scandinavian Journal of Work, Environment and Health*, *32*(5), 402–412.

Martínez, K., Burgaleta, M., Román, F. J., Escorial, S., Shih, P. C., Quiroga, Mª. A., et al. (2011). Can fluid intelligence be reduced to 'simple' short-term storage? *Intelligence*, *39*, 473–480.

Martínez, K., Solana, A. B., Burgaleta, M., Hernández-Tamames, J. A., Álvarez-Linera, J., Román, F. J., et al. (2013). Changes in resting-state functionally connected parieto-frontal networks after videogame practice. *Human Brain Mapping* (in press).

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

Melby-Lervåg, M., & Hulme, C. (2012). Is working memory training effective? A meta-analytic review. *Developmental Psychology*. http://dx.doi.org/10.1037/a0028228 (Advance online publication).

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

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.)Los Angeles, CA: Muthén & Muthén.

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

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., et al. (2012). Intelligence: New findings and theoretical developments. *American Psychologist*. http://dx.doi.org/10.1037/a0026699 (Advance online publication).

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

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. *Neuropsychologia*, *9*, 97–113.

Posner, M. I., & Rothbart, M. K. (2007). *Educating the human brain.* Washington, DC: American Psychological Association.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*, 193–203.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., et al. (2012). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*. http://dx.doi.org/10.1037/a0029082 (Advance online publication).

Reeve, B. B., Hays, R. D., Bjorner, J. B., et al. (2009). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(Suppl. 1), 22–31.

Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence*, *33*, 535–549.

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, *84*(3), 228–238.

Rudebeck, S. R., Bor, D., Ormond, A., O'Reilly, J. X., & Lee, A. C. (2012). A potential spatial working memory training task to improve both episodic memory and fluid intelligence. *PLoS One*, *7*(11), e50431. http://dx.doi.org/10.1371/journal.pone.0050431.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* Monograph, No: Psychometrika 17.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*. http://dx.doi.org/10.1037/a0027473 (Advance online publication).

Stephenson, C. L., & Halpern, D. F. (2013). Improved matrix reasoning is limited to training on tasks with a visuospatial component. *Intelligence*, *41*, 341–357.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 331–352.

te Nijenhuis, J., van Vianen, A. E. M., & van der Flier, H. (2007). Score gains on *g* loaded tests: No *g*. *Intelligence*, *35*, 283–300.

Thissen, D., & Wainer, H. (2001). *Test scoring.* Mahwah: Erlbaum.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., et al. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402.

von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language*. http://dx.doi.org/10.1016/j.jml.2013.02.002.

Wickstrom, G., & Bendix, T. (2000). The "Hawthorne effect" — What did the original Hawthorne studies actually show? *Scandinavian Journal of Work, Environment and Health*, *26*(4), 363–367.

World Medical Association (2008). Declaration of Helsinki — Ethical principles for medical research involving human subjects. *59th WMA General Assembly, Seoul, Korea*.