Intelligence 41 (2013) 341–357

Contents lists available at SciVerse ScienceDirect

Intelligence

journal homepage:

Improved matrix reasoning is limited to training on tasks with a visuospatial component ☆

Clayton L. Stephenson [a,*], Diane F. Halpern [b]

[a] School of Behavioral and Organizational Sciences, Claremont Graduate University, United States
[b] Department of Psychology, Claremont McKenna College, United States

ARTICLE INFO

Article history:
Received 17 January 2012
Received in revised form 6 May 2013
Accepted 24 May 2013
Available online xxxx

Keywords:
Fluid intelligence
Dual *n*-back task
Working memory capacity

ABSTRACT

Recent studies (e.g., Jaeggi et al., 2008, 2010) have provided evidence that scores on tests of fluid intelligence can be improved by having participants complete a four week training program using the dual *n*-back task. The dual *n*-back task is a working memory task that presents auditory and visual stimuli simultaneously. The primary goal of our study was to determine whether a visuospatial component is required in the training program for participants to experience gains in tests of fluid intelligence. We had participants complete variations of the dual *n*-back task or a short-term memory task as training. Participants were assessed with four tests of fluid intelligence and four cognitive tests. We were successful in corroborating Jaeggi et al.'s results, however, improvements in scores were observed on only two out of four tests of fluid intelligence for participants who completed the dual *n*-back task, the visual *n*-back task, or a short-term memory task training program. Our results raise the issue of whether the tests measure the construct of fluid intelligence exclusively, or whether they may be sensitive to other factors. The findings are discussed in terms of implications for conceptualizing and assessing fluid intelligence.

Published by Elsevier Inc.

1. Introduction

Working memory (WM) is the ability to maintain information in memory while performing one or more tasks (Miyake & Shah, 1999). WM is a vital cognitive function that may not be domain specific and is important in complex cognitive abilities such as reasoning (Kyllonen & Christal, 1990), reading comprehension (Daneman & Carpenter, 1980), writing (McCutchen, 1996), note taking (Kiewra & Benton, 1988), and fluid intelligence (G*f*; Ackerman, Beier, & Boyle, 2005). Although many tasks (e.g., writing and note taking) become easier and more automatic with practice, increasing performance for one type of task (e.g., note taking) does not necessarily mean that performance in a different type of task (e.g., reasoning) will be improved. But, what if WM, which is not domain specific and is necessary for completing a number of cognitive tasks, could be improved? Would improvements in WM transfer to a wide range of tasks? To answer this question, we focused on the effects of training to improve WMC on tests of G*f*, visuospatial abilities, and verbal abilities.

1.1. G*f*'s relation to WM

Conceptually, G*f* is the ability to solve novel problems that are not solvable using previously learned strategies (Carroll, 1993). More specifically, G*f* is an indication of a person's general ability for sequential and inductive reasoning, performance levels on Piagetian reasoning tasks, and processing speed for reasoning (McGrew, 1997). To solve novel problems or develop new types of reasoning, a person usually needs to hold relevant information in temporary storage

☆ The data in the current paper was part of Clayton Stephenson's doctoral dissertation. We thank Dale E. Berger, Gabriel I. Cook, and James C. Kaufman for providing valuable feedback on the dissertation and the current paper.

* Corresponding author at: Department of Psychology, University of Southern California, SGM 501, 3620 South McClintock Ave., Los Angeles, CA, 90089-1061, United States.
E-mail address: clstephe@usc.edu (C.L. Stephenson).

0160-2896/$ – see front matter. Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.intell.2013.05.006

while simultaneously manipulating and processing incoming information, which is the accepted definition of working memory (Heitz, Unsworth, & Engle, 2004). Working memory capacity (WMC) refers to the number of units (i.e., amount of information) that can be stored while processing incoming information and ignoring irrelevant information and, is often referred to as the control mechanism in WM (Cowan, 2010).

Using structural equation modeling (SEM), Colom, Abad, Rebello, and Shih (2005) showed that WM and *g* are almost isomorphic, but only when WM is measured using tasks that requires processing *and* storage of the information. However, Colom, Rebollo, Abad, and Shih (2006) also showed through SEM procedures that short-term memory (STM) was a better predictor of G*f*, crystallized intelligence (G*c*), perceptual speed, and spatial relations than WM on six out of eight interpretable models and that STM and WM were equally strong predictors of G*f* in two of the eight interpretable models. Because the measures of STM and WM are not distinguishable, Colom et al. (2006) suggested that the common factor is not attentional control as suggested by Engle, Tuholski, Laughlin, and Conway (1999) and Kane et al. (2004), but that the common component is a reliable short-term storage of information. Similar findings that the short-term storage capacity is the underlying process linked to G*f* have been replicated by Martínez et al. (2011); Krumm, Schmidt-Atzert, Buehner, Ziegler, Michalczyk, and Arrow (2009); and includes studies measuring WM and G*f* in children (Hornung, Brunner, Reuter, & Martin, 2011).

Based on the previous studies using SEM, it is apparent that a reliable short-term storage is an important cognitive component when G*f* is being tested because test items for G*f* require the test taker to reliably remember rules and patterns while processing each component of the problem. With a reliable and successful short-term storage component, people are able to develop a solution while ignoring other patterns or distractions embedded in the problem. The Raven Progressive Matrices (Raven, 2000) are tests that require test takers to use such cognitive processes and, are the most accepted and frequently used measure of G*f*. The Raven's Matrices consist of figures that are organized in 3×3 matrices. The figures vary systematically across rows and across columns, though the last object in the sequence (i.e., lower right-hand corner) is left blank. The test taker's task is to choose from several alternatives the one figure that correctly completes the sequencing. WMC is important because the test taker must recognize patterns in the matrices and store the relevant information while ignoring other patterns that are irrelevant to the solution.

Like other measures of intelligence, the measurement of G*f* is not immune to controversy. The Raven's tests were designed to measure the unidimensional construct, G*f* (Raven, 2000; van der Ven & Ellis, 2000). Schweizer, Goldhammer, Rauch, and Moosbrugger (2007) partially supported the assumption that the Raven's Advanced Progressive Matrices (APM) is unidimensional by testing its convergent validity (i.e., how well similar constructs converge onto a single factor) and discriminant validity (i.e., how well constructs that are not theoretically related fail to load onto the same factor) using Horn's (1983) reasoning and visualization scales as the comparison. Schweizer et al. found significant convergent validity ($r = .68$) between the scores on the APM and Horn's reasoning scale. However, Schweizer et al. also found weaker correlations (r ranging from .24 to .34) between the APM and Horn's visualization scale for the dissimilar items. Overall, the APM's correlation with reasoning is stronger than the correlation with spatial ability, thus, supporting the idea that the APM is a relatively pure measure of G*f*.

Despite the statistically strong convergent validity suggesting that reasoning and the ability to solve matrix problems as measured by the APM, the poor discriminant validity between spatial abilities and the APM has implications for the relationship between visuospatial abilities and G*f*. A corollary of the hypothesis that visuospatial abilities underlie success in the APM is that if people can improve their visuospatial abilities, their scores on the APM would also increase, but this improvement would be a reflection of improved visuospatial abilities and not an improvement in G*f*. Many studies have found that playing action video games can improve visuospatial abilities such as mental rotation (Cherney, 2008; Feng, Spence, & Pratt, 2007; Okagaki & Frensch, 1994), spatial attention (Feng et al., 2007), spatial visualization (Okagaki & Frensch, 1994), spatial resolution of visual processing (Green & Bavelier, 2007), visuospatial attention (Castel, Pratt, & Drummond, 2005; Green & Bavelier, 2003; Green & Bavelier, 2006a), and the ability to track multiple objects at once (Green & Bavelier, 2006b). Even if playing video games in an everyday context influenced a wide range of visual–spatial laboratory tasks, the effects may not generalize to visuospatial tasks in the real world or to any cognitive ability, such as G*f*. It may be that practice at tasks that require the use of a domain-general cognitive abilities such as working memory capacity (WMC) are needed for transfer to other domain-general cognitive abilities such as G*f*.

1.2. Evidence for improving Gf

Recent research by Jaeggi, Buschkuehl, Jonides, and Perrig (2008) and Jaeggi et al. (2010) has provided evidence that a cognitive training program designed to enhance working memory capacity (WMC) improved scores on the APM and the Bochumer Matrices Test (BOMAT), which is a recently created test that is similar to the APM, but more difficult, thus showing improvements in G*f*. Jaeggi et al.'s, 2008 study had participants complete a training program using the dual *n*-back task as the training. The dual *n*-back task is a complex task that requires participants to simultaneously store in memory and process visual and auditory stimuli (see Method for a more detailed description). The four-week training program was designed with five training sessions per week with each training session lasting approximately 20 min. Participants' scores on the tests of G*f* improved (Cohen's $d = .65$) after they completed the dual *n*-back training program over spans of 8, 12, 17, and 19 days. These scores were compared to a no-training control group. Jaeggi et al.'s, 2010 study was similar in method and results to their 2008 study, but included an additional group that trained using the visual *n*-back, which is similar to the dual *n*-back task, but presents only visual stimuli. Participants in Jaeggi et al.'s (2010) study who completed the visual *n*-back training also experienced gains in scores on tests of G*f* ($d = .65$) just as the participants who completed the dual *n*-back training ($d = .98$).

A question raised from Jaeggi et al.'s (2008, 2010) findings is, "How were Jaeggi et al. able to enhance G*f* with a relatively short training program?" Part of the answer is that Jaeggi et

al. used a dual processing task with the intention of improving WMC, which has a strong influence on G*f*. However, visuospatial abilities are also important in completing tests of G*f*. By using the dual *n*-back task and the visual *n*-back task as training for WMC, there is no definitive indication of which mechanisms were enhanced in Jaeggi et al.'s studies.

Another important element of Jaeggi et al.'s (2008, 2010) studies is that there were no statistical comparisons made to determine differences between men and women regarding gains on the APM. It is important to compare men and women on tests of G*f* because there are well-documented sex differences in visuospatial abilities. For example, Linn and Petersen (1985) found a large effect size – using Cohen's *d* – of sex differences for mental rotation (.73), a moderate effect size for spatial perception (.44), and a small effect size for spatial visualization (.13), all favoring men. Voyer, Voyer, and Bryden (1995) also found the largest effect size in mental rotation at .66, followed by spatial perception at .48, and an effect size of .23 for spatial visualization. Hyde (2005) conducted a meta-analysis of other meta-analyses and also found that the largest difference that exists between men and women regarding cognitive abilities was in the visuospatial abilities category with Cohen's *d* ranging from .13 to .73. Overall, based on meta-analyses spanning two decades, there is a robust effect size in sex differences in visuospatial abilities, especially for people older than 18 (Linn & Petersen, 1985). It could be that because men already have an advantage in some visuospatial abilities, they may not experience the same magnitude of increased scores on tests of G*f* as women when G*f* is measured with visual–spatial tasks.

A number of theorists (e.g., Jensen, 1998) have viewed G*f* as being a biologically predetermined ability. The results of Jaeggi et al.'s (2008, 2010) studies, however, have significant implications for the way philosophers, psychologists, and educators think about intelligence because the take-home message is that the ability to solve novel problems can be improved with a short training program. Their studies also have implications for the psychometric properties and uses of the APM and other tests of G*f*. Therefore, a careful analysis is needed to substantiate and determine WMC mechanisms that are being improved and leads to improvement in G*f*. The current study sought to determine what mechanisms in WMC might be improved through cognitive training, whether there are sex differences in the improvements, and determine the generalizability across other measures of G*f* and cognitive tests.

2. Method

2.1. Participants and design

One hundred thirty-nine participants (73 women, 66 men) were recruited from a state university and a private and highly selective liberal arts college in California. One male and two women were excluded from the data analysis resulting in a total of 136 participants (71 women, 65 men).[1] One hundred thirty-four participants received course credit or extra credit for participating in the study while the other five participants participated strictly as volunteers. All participants were also entered into a raffle to win a $15 gift card. The raffle occurred once per week for each training group. A total of 110 participants completed the training sessions; 26 participants were in the control group.

A primary concern in the current study was recruiting a representative sample that was equally distributed across all conditions. Participants mean age was 22.48 (5.83) with no difference between men and women, $t\ (134) = -.53$, $p = .60$, and no statistically significant difference among training groups, $F\ (4, 131) = 2.11$, $p = .08$. Participants' education was also calculated such that education included kindergarten through twelfth grade (i.e., 13 years) plus years of college completed. Participants mean education was 15.57 (2.23) years with no difference between men and women, $t\ (134) = .29$, $p = .77$, and no difference among training groups, $F\,(4, 131) = .63$, $p = .64$.

Participants were randomly assigned to training conditions regardless of the school in which they were enrolled. A loglinear test was conducted to determine if participants were equally distributed based on sex and the university in which students were enrolled. There was no significant difference in the number of women and men from each school per training condition, $\chi^2\ (13) = 2.99$, $p = .99$. There were also no differences in the distribution of participants to training groups based on the university in which they were enrolled, $\chi^2\ (4) = 1.39$, $p = .85$. Overall, participants were equally distributed among all conditions.

A 2 (participants' sex; women and men) × 2 (repeated testing; pretest and posttest) × 5 (type of training; dual *n*-back, visual *n*-back, auditory *n*-back, spatial matrix span, and control) repeated measures mixed design was implemented. Participants were randomly assigned to one of five training groups. Four of the groups received one of the following types of cognitive training: a) the dual *n*-back task (14 men, 14 women), b) the visual *n*-back task (14 men, 15 women), c) the auditory *n*-back task (12 men, 13 women), or d) the spatial matrix span (13 men, 15 women). The fifth group received no training and acted as a control group (12 men, 14 women). In accordance with Jaeggi et al.'s (2008) design, participants trained five days a week (Monday through Friday) for four weeks, for approximately 20 min each day. Participants were also divided into groups based on sex with approximately equal numbers of men and women in each group. For repeated testing, all participants' G*f* and cognitive abilities were assessed two times: prior to beginning the training (pretest) and after completing the training (posttest), with the no training control group assessed at the same time. The posttest was taken within four days of completing the final training session.

2.2. Materials

2.2.1. Training session tasks

2.2.1.1. Dual n-back task. The dual *n*-back task is a complex task with multiple elements (see Fig. 1 for a visual representation). First, the letter "n" in the term "*n*-back" refers to the number of trials back that a participant must remember that the target was presented. For example, a *2*-back task requires the

[1] The male was excluded from the auditory *n*-back task condition because his daily average training performance levels were double the group's average. The two women did not go beyond *2*-back throughout the entire dual *n*-back task training program even though they were instructed each day on how to complete the task successfully, thus, they were excluded from the analysis.

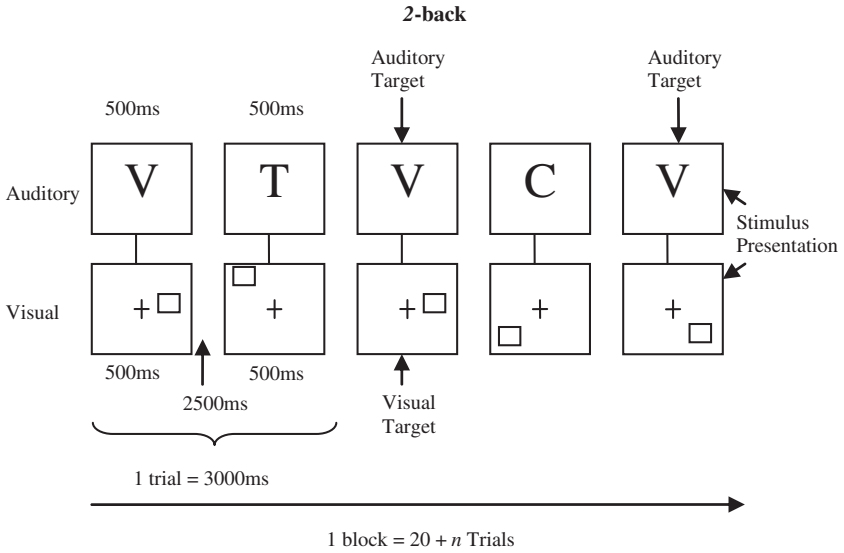


Fig. 1. Display of five trials on the dual *n*-back task.

participant to remember whether a target stimulus was presented two trials prior to the current stimulus presentation. Thus, participants must continually update their memory to accommodate a new stimulus as a potential item to be remembered in the future.

The second important element to the *n*-back task is the presentation of stimuli. Participants were first presented with a white cross as a fixation point on a dark background for 2500 ms. Participants then simultaneously saw a blue box appear in one of eight positions on the computer screen (i.e., top left, top center, top right, middle left, middle right, bottom left, bottom center, or bottom right) and heard one of eight consonants (i.e., c, d, g, k, p, q, t, and v) spoken into headphones. Although it seems that the consonants could be easily confused because of their similarity, Jaeggi et al. (2007, 2008) did not report participants having any difficulty distinguishing among them. The visual and auditory stimuli were presented for 500 ms. The presentation of the stimuli was followed by an inter-stimulus interval (ISI) that showed only the white cross against a dark background, which lasted for 2500 ms. During the 2500 ms ISI, the participant was supposed to press the letter "A" on the keyboard to indicate the blue square was presented *n*-back trials or to press "L" to indicate the consonant was presented *n*-back trials. A trial represents one presentation of a stimulus (500 ms) *and* one ISI (2500 ms); thus, one trial equals 3000 ms.

The third element to understanding the *n*-back task is that a block is a set of 20 + *n* trials. The *n* indicates that additional trials may be needed depending on the number of *n*-backs that a participant was asked to remember. For example, if the block was a three *n*-back task, then three additional trials were needed. The challenging part of the dual *n*-back task is that it requires participants to constantly update their memory storage for two modalities at the same time in order to complete the task successfully.

The dual *n*-back task used by Jaeggi et al. (2008) varied in difficulty for each block depending on the participant's performance. Each training session began with a *2*-back task. If the participant made more than five mistakes, then the next block would be decreased (e.g., to a *1*-back task). If the participant made fewer than three mistakes, then the next block would be increased (e.g., to a *3*-back task). If the participant made three to five mistakes, then the next block would remain at the same level (e.g., a *2*-back task). Jaeggi et al. found that, on average, participants make it to *6*-back on the nineteenth day of training.

All of the cognitive tasks used for training changed in difficulty based on the participant's performance for each block. The adapting feature of the dual *n*-back training was the same for every training regimen and participants in all training groups trained for a fixed 20-minute session. If the participant was in the middle of a block when the 20 min had passed, the participant was able to complete the block and then the program stopped the training. All stimulus materials were developed by Jaeggi and colleagues.

2.2.1.2. Visual n-back task. The visual *n*-back task was used as an alternative WMC training regimen. This task was designed to test whether the visuospatial component of the *n*-back task was responsible for the increase in scores on tests of G*f* in Jaeggi et al.'s (2008) study. The visual *n*-back task is similar to the dual *n*-back task with the exception that only the visual stimuli (i.e., the blue boxes) are presented. The goal for participants was the same as the dual *n*-back task in that they had to remember whether the target stimulus was presented in the previous one, two, three, or *n*-back trials. The visual *n*-back task adapted to the participants' performance using the same rules as the dual *n*-back task.

2.2.1.3. Auditory n-back task. The auditory *n*-back task was also used as an alternative for WMC training regimen. The auditory *n*-back task was used to further address two possible confounds in Jaeggi et al.'s (2008, 2010) studies. First, it was used to determine if the visuospatial component of the dual *n*-back task was necessary for the improvement in participants' scores on tests of G*f*. Second, it was also used to

determine if the nature of a working memory task was responsible for the increase in scores without the need for visuospatial stimuli. The auditory *n*-back task is similar to the dual *n*-back task in that it uses the same consonants. It differs in that it does not present any visual stimuli. The task of the participants was the same as the visual *n*-back task in that the target stimulus was determined as being presented one, two, three, or *n*-back trials. This task also changed in difficulty for each block based on participants' performance.

2.2.1.4. Spatial matrix span. The spatial matrix span task was given to participants as a control for the WMC training regimens. This task was used in Kane et al. (2004) and is considered to be a spatial STM task. The spatial matrix span is much like the game Concentration and begins with a presentation of a 4 × 4 matrix on the computer screen. A blue circle was used as the stimulus and was presented in one of the sixteen cells at a time (see Fig. 2). Each blue circle was presented for 500 ms. The order of cells in which the blue circle was presented was random. For example, Fig. 2 shows the blue circle appearing in the cell where the third column intersects the second row in the first trial. This is followed by the blue circle appearing in a cell where the second column intersects the fourth row in the second trial. The participants' task was to recall the sequence of blue circles appearing in the cells by moving the cursor and clicking the mouse in the appropriate cells in the correct order. The spatial matrix span training adapted to participant's performance, but on a trial to trial basis rather than after 20 or more trials as the case of the *n*-back training programs. The task started with having the participant remember two positions within the matrix. If the participant remembered the sequence correctly, then the number of sequenced positions increased by one sequence on the following trial. If the participant made a mistake, then the number of positions to be remembered decreased by one. Similar to the other tasks, the spatial matrix span lasted 20 min, but the training continued past 20 min if the participant was in the middle of completing a block.

2.2.2. Measures of Gf

Four measures were used to assess participants' G*f*: Raven's APM, Cattell's Culture Fair Test, the Matrix Reasoning subtest for Wechsler's Abbreviated Scale of Intelligence (WASI), and the Matrix Reasoning subtest for the BETA-III. The tests of G*f* were chosen because of their sound psychometric properties and because they have recently been updated in their standardizations and norms (Kellogg & Morton, 1999; Raven, 2000; WASI Manual, 1999). These tests were also used by Engle et al. (1999) and Kane et al. (2004) to load on G*f* as a latent exogenous variable in their SEM procedures. More specifically, Engle et al. used the APM and Cattell's Culture Fair Test to measure G*f*. Kane et al. used the APM, the WASI Matrix Reasoning subtest, and the BETA-III Matrix Reasoning subtest as measures of G*f*. Participants were given practice items for each test before taking the actual test.

A time restriction was implemented on all measures of G*f*. The first and foremost reason for collecting response time measures was that Jaeggi et al. (2008, 2010) restricted time in their studies. The implementation of time restrictions on the APM does not reduce the reliability or validity. Jaeggi et al. stated that the correlation between timed and untimed versions of the APM is .95 and Raven, Raven, and Court (1998) stated that intellectual efficiency is more likely to be assessed when participants are timed. Second, the Cattell's Culture Fair Test and the BETA-III Matrix Reasoning subtest were originally designed for time-limited administration. The WASI Matrix Reasoning subtest is not typically timed when used with the rest of its subtests. However, a five minute time restriction on the WASI was used because it is a similar test to the BETA-III in difficulty and the number of items in each test (i.e., WASI has 29 problems and BETA-III has 25 problems). In general, all measures were administered with a time limit to maintain consistency in procedures across all tests.

All tests were administered in their entirety for the pretest and posttest. Practice effects, of course, are an issue, but were accounted for by having a control group that did not have training. If there were improvements due to practice effects exclusively, then the control group should experience gains comparable to the training groups. However, training groups should have higher scores than the control group if the training actually works above and beyond practice effects.

2.2.2.1. Raven's Advanced Progressive Matrices. The APM assesses a person's ability to develop a solution for a problem by reasoning inductively (Wilhoit & McCallum, 2003). Items in the APM test progress in their difficulty. Eight black and white items are presented within the box in a 3 × 3 matrix with an empty space for the ninth cell. Eight choices, each with a different item, are provided below the box. The participants' task was to choose one of eight possibilities to complete the sequence. Jaeggi et al. (2008, 2010) and Kane et al. (2004) used

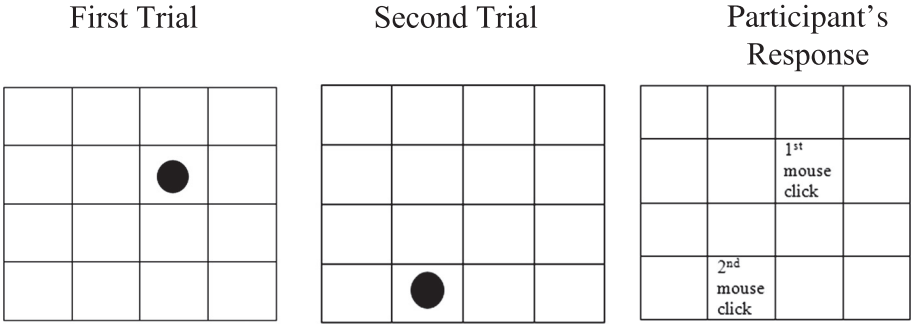


Fig. 2. Sequence of events for the Spatial Matrix Span training. Circles are blue in the actual training. The words in the matrix labeled "Participant's Response" do not appear during the test. The words are simply indicating what the participant must do in order to complete the task correctly.

Set II of the APM and so did the current study. The current study implemented the same time restriction as Jaeggi et al. (2008) by allowing participants to solve as many of the 36 problems as they could in only 10 min. The APM was not divided into parallel forms for the pretest and posttest because we wanted to assess potential benefits participants would experience when taking the test as they might in a real-world setting (e.g., taking the test once during a job application process and having to take it again in a separate job application process).

2.2.2.2. Cattell's Culture Fair Test. The 1963 version of Cattell's Culture Fair Test, Scale 3, Form A, was used as a measure of G*f*. Cattell's Culture Fair Test contains four problem subtests that are timed. Each subtest is taken with a time restriction and each subtest progresses in difficulty. The first problem subtest contains 13 items that test the ability to complete a series that shows a progressive change in an object, shape, or figure. The second problem subtest contains 14 items that test the ability to classify objects by identifying two objects, shapes, or figures that do not belong in a set of five. The third problem subtest test contains 13 items that test the ability to complete a matrix with four to nine boxes in each matrix. One block of the matrix is missing and the participants' task is to choose the best figure from six choices to complete the matrix. The fourth problem subtest contains ten items with dots, lines, and various geometrical shapes in each item. Each item contains a cell on the left hand side that shows the relationship between the lines, dots, and shapes. The relationship of each item is based on rules that determine the placement of the dot among the other shapes and lines. The participants' task is to choose one item of five that shows the same relationship between the dot and the objects. In order to complete this task, the participants must 1) identify the rules governing the relationship between the dot and objects and 2) imagine the dot's position in the five choices to determine which one fits the same rules.

2.2.2.3. WASI Matrix Reasoning subtest. The Matrix Reasoning subtest is one of four tests in the WASI, which is an abbreviated test of the Weschler Adult Intelligence Scale (WAIS). Only the Matrix Reasoning subtest was given to participants because it is the portion of the test that assesses G*f*. The items are presented in a matrix with all but one of the cells containing colored figures. Below the matrix are five boxes that each contains a figure. Similar to the APM, the WASI Matrix Reasoning subtest presents figures in a matrix where each figure either remains constant or changes in each cell. The participants' task is to choose the answer from one of the five boxes to complete the sequence. A time restriction was implemented such that participants had 5 min to complete 29 items.

2.2.2.4. BETA-III Matrix Reasoning subtest. The Matrix Reasoning subtest is one of five tests of the BETA-III. Only the Matrix Reasoning subtest (i.e., Test 5) was given to participants. BETA-III problems are presented in a 2×2 matrix that contain black and white figures and are very similar to WASI Matrix Reasoning subtest; thus, no figure is reproduced. Just like the WASI Matrix Reasoning subtest, each matrix in the BETA-III shows figures in a sequence with an empty cell. To the right hand side of the matrix are five cells that each contains a different figure. The participants' task is to choose the cell that contains the figure that will complete the sequence. The test contains 25 questions and, based on the test's instructions, participants were given 5 min to complete all questions.

2.2.3. Cognitive assessments

Four cognitive tests were used to determine if the WMC training transfers beyond a psychometric test of G*f*. Because tests of G*f* are thought to have visuospatial components (van der Ven & Ellis, 2000), the Mental Rotation Test (Shepard & Metzler, 1971; Vandenburg & Kuse, 1978) and the Paper Folding Test (Ekstrom, French, Harman, & Dermen, 1976) were used to test visuospatial abilities. If the dual *n*-back task is improving visuospatial abilities, then there should be an improvement of performance on the two visuospatial tests. These two visuospatial abilities tests were used because the Mental Rotation Test taps into mental rotation abilities and the Paper Folding Test taps into spatial visualization abilities (Linn & Petersen, 1985). The Extended Range Vocabulary Test (Ekstrom et al., 1976) and the Lexical Decision Test (Ratcliff, Gomez, & McKoon, 2004) were used because they are tests of verbal abilities and a goal of the current study is to determine if WMC training will transfer to tasks that are not visuospatial. All cognitive tests were presented on a 15.6 in. laptop screen using SuperLab 4.0 and were given during the pretest and posttest sessions. The cognitive tests were not administered with time restrictions that are typically used (e.g., three minute time restriction for the Paper Folding Test) because reaction times were recorded to determine if participants' response times would decrease, in addition to increasing in number correct, as a result of training. Each test is discussed separately.

2.2.3.1. Mental Rotation Test. Vandenburg and Kuse's (1978) Mental Rotation Test was used as a test of participants' ability to mentally rotate an object. Twenty items from the paper-and-pencil version were scanned into a computer and presented in SuperLab 4.0 to record response times. The stimuli for the Mental Rotation Test represent three-dimensional objects made of ten small blocks. Three different objects are presented on the right hand side of the computer screen and it is the task of the participants to match one of three objects with a target object (presented on the left hand side of the computer screen). All 20 items were presented in the pretest and the posttest.

2.2.3.2. Paper Folding Test. The Paper Folding Test was adapted from the Educational Testing Service kit (Ekstrom et al., 1976). Similar to the Mental Rotation Test, 18 stimuli were scanned into a computer and presented on SuperLab 4.0 so RTs could be measured. The Paper Folding Test stimuli consist of a step-by-step presentation of a paper being folded with a minimum of one fold to a maximum of three folds. The final step shows a hole punched in the paper. The task of the participant is to choose from five choices what the punched paper would look like unfolded. All 18 stimuli were presented in the pretest and posttest.

2.2.3.3. Extended Range Vocabulary Test. The Extended Range Vocabulary Test is another test from the Education Testing Service kit (Ekstrom et al., 1976). This test was also presented

on the computer so that we could collect response times. It is a standard vocabulary test with 48 words. The test is presented such that a target word is given in the middle of the screen. Five different words are presented below the target word and the participants' task is to choose which word best defines the target word by pressing a number on the keyboard that is associated with that word. For example, the target word might be "bantam." The choices for this target are, "1) fowl, 2) ridicule, 3) cripple, 4) vegetable, and 5) ensign." The correct answer is "fowl." All 48 words were presented in the pretest and posttest.

2.2.3.4. Lexical Decision Test. The Lexical Decision Test (Ratcliff et al., 2004) uses a presentation of nonwords and words. All nonwords were selected using the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002). Twenty nonwords were selected to be monosyllable and be pronounceable nonwords such as "yumph." All words were selected using the MRC Psycholinguistic Database (Coltheart, 1981). Twenty monosyllable, high-frequency English words such as "staff" were selected randomly. Words and nonwords are presented in random order on a computer screen. The participant's task is to indicate if the target stimulus is a real word or a nonword by pressing one of two different keys on a keyboard. Rather than being a measure of the participants' understanding of the meaning of the words, which is the purpose of the Extended Range Vocabulary Test, the Lexical Decision Test measures the speed of a participants' access to verbal information to recognize a word based on its orthography (Halpern & Wai, 2007). All 40 items were presented in the pretest and posttest.

2.3. Procedure

2.3.1. Pretest session

Participants were given the tests of G*f* and the cognitive tests one week prior to the first training session. Participants were seated in a well lit room that was free of distractions. They were randomly assigned to complete either the tests of G*f* or the cognitive tests first. To avoid any practice effects or other systematic confounds, the order of the tests of G*f* and the cognitive tests were also randomized for each participant. Participants were allowed 10 min to complete the APM, approximately 13 min to complete Cattell's Culture Fair Test, 5 min to complete the WASI Matrix Reasoning subtest, and 5 min to complete the BETA-III Matrix Reasoning subtest. The total time to take the tests of G*f* was approximately 55 to 60 min, including time to give instructions, answer any questions, and actual testing time. The cognitive tests took approximately 15 to 30 min to complete, including instructions and time to answer questions. The total time to complete all tests ranged from 70 to 90 min.

2.3.2. Training sessions

Participants were seated in front of a computer monitor, keyboard, and mouse in a room that was well lit and free of distractions. Participants who were assigned the dual *n*-back task or the auditory *n*-back task wore headphones to hear the consonants being spoken. All training sessions were presented with BrainTwister 1.0.2, a program designed by Buschkuehl, Jaeggi, Kobel, and Perrig (2008). For the first two days of training, the experimenter provided verbal instructions for the participants in addition to the written and visual instructions to make sure the participant understood the task. The training stimuli began after the participant pressed the space bar on the keyboard.

Each participant completed 18 to 20 training sessions and each session lasted approximately 20 to 22 min, which is slightly shorter than Jaeggi et al.'s (2008) sessions that lasted approximately 25 min. The time for each session varied slightly so that if the 20 minute mark arrived and the participant was in the middle of a block, then the participant was allowed to complete the block. Participants were allowed to take breaks, which resulted in some participants taking 30 min to complete the training in a session. The computer program indicated the completion of the daily session to the participant. Attrition and missed training sessions were a concern; therefore, strict criteria were set in case a participant missed a training session. First, if participants did not show up for the first two training sessions, then they were not allowed to continue the study. Second, if participants missed a training session, then they were allowed to make up the training session on the following day by completing two training sessions. However, participants were not allowed to make up more than two training sessions. Only 5% of the participants who completed training did not have to make up a training session.

2.3.3. Posttest session

The same tests of G*f* and cognitive tests used in the pretest session were used in the posttests. The same testing conditions and random ordering procedures were also used. The posttest was given no later than four days after the last training interval.

2.4. Hypotheses

The additions to Jaeggi et al.'s (2008, 2010) studies allowed us to determine 1) if training used to improve scores on tests of G*f* requires a visuospatial component; 2) if the effects of training WMC transfer to other tests of G*f* aside from the APM and the BOMAT; 3) if women experience greater gains than men in scores on tests of G*f* after training; and 4) if the effects of training transfer beyond tests of G*f* to more domain specific tasks (e.g., to other visual–spatial tasks such as the Mental Rotation Test). To address the first issue, we incorporated an auditory *n*-back task and a spatial matrix span task as separate training conditions. We predicted that participants assigned to train WMC using the auditory *n*-back task or train STM using a spatial matrix span task would have little or no improvement on scores of tests of G*f* when compared to participants who completed training using the dual *n*-back or visual *n*-back task. However, because the auditory *n*-back is still considered to be a WMC task, there is the potential that scores on tests of G*f* may still be improved. Therefore, we predicted improvements on tests of G*f* for participants who completed the auditory *n*-back task during training compared to participants who completed training using a spatial matrix span or the no-training control group.

The issue of transfer to other measures of Gf was addressed by using multiple measures of G*f* including the APM, the Wechsler Abbreviated Scale of Intelligence (WASI) Matrix Reasoning subset, Cattell's Culture Fair Test, and the BETA III Matrix Reasoning subset. If these four tests of G*f* are measuring the same construct, then participants who train with the dual

n-back or visual *n*-back training should experience significant improvements on all four of the tests. However, participants who complete training using the auditory *n*-back task or the spatial matrix span or who were in the no training group should experience little or no improvement. We predict that participants who complete auditory *n*-back training should show little improvement in their scores on tests of G*f* because the auditory *n*-back does not have a visuospatial component. The spatial matrix span is a STM task that does not contain a dual component (i.e., maintaining information in memory while processing other information for a separate task), thus, does not tax WMC the same way other tasks do.

The issue regarding potential sex differences in gains on tests of G*f* was addressed by recruiting near-equal numbers of men and women for each training program. Women and men who train using the dual and visual *n*-back tasks are both expected to experience gains on G*f*; however, women are expected to make greater gains than men if the improvement is largely due to improved visuospatial abilities (see Feng et al., 2007). If WMC is the cognitive function that is being improved through training, then the men and women in the dual *n*-back group were expected to experience similar gains on all four tests of G*f*. If visuospatial ability is the cognitive function that is being improved, then we expected women in the visual *n*-back group to experience greater gains than men. Neither men nor women in the auditory *n*-back training group, spatial matrix span, or control group were expected to experience any gains on tests of G*f*.

The issue concerning transfer of training beyond tests of G*f* was addressed by giving participants four cognitive tests during the pretest and the posttest sessions. The four cognitive tests included the Mental Rotation Test and the Paper Folding Test as visuospatial tests and the Lexical Decision Test and an Extended Range Vocabulary Test as the verbal tests. The cognitive tests were used to test if training transfers to the domain-specific tasks of visuospatial and verbal abilities. Visuospatial and verbal skills are, for the most part, independent of each other (Halpern, 2012). Therefore, improvement in visuospatial skills should not transfer to verbal skills. Based on this reasoning, verbal abilities were tested to ensure domain specific improvement instead of improvements being made as a result of familiarity with visuospatial tasks (Wright, Thompson, Ganis, Newcombe, & Kosslyn, 2008). Visuospatial and verbal abilities are quite different from one another and improvement in one domain should not transfer to another. Thus, the predictions for improvements on cognitive tasks are as follows. First, participants who complete training using the dual *n*-back or visual *n*-back task will experience improvements in reaction time and number correct on the Mental Rotation Test and Paper Folding Test compared to the participants in the auditory *n*-back, spatial matrix span, or the control condition. Second, no improvements are predicted for the Lexical Decision Test and Extended Range Vocabulary Test for any of the training conditions. Third, women who train using the dual or visual *n*-back tasks are expected to make greater gains than men on the Mental Rotation Test and the Paper Folding Test.

We had two primary goals for the current study: 1) determine the factors in cognitive training that lend to improvements in scores on tests of *Gf*, and 2) determine how generalizable Jaeggi et al.'s (2008, 2010) results were in men and women, tests of *Gf*, and tests of cognitive abilities. Overall, we predicted that participants, especially women, who train WMC using the dual *n*-back task or the visual *n*-back task will experience the greatest improvements on tests of G*f*, the Mental Rotation Test, and the Paper Folding Test when compared to participants who complete training using the auditory *n*-back task, the spatial matrix span, or the control group. However, participants who complete training using the auditory *n*-back task will experience greater improvements on tests of G*f* when compared to participants who complete training using the spatial matrix span or the control group.

3. Results

The pretest data for tests of G*f* and cognitive tests were first analyzed to assure that there were no significant differences in performance among training groups resulting in unintentional biases. Mean scores for pretests and posttests of G*f* were calculated based on the number of correct answers in each test (see Table 1 for means and standard deviations). A MANOVA was conducted to determine if there were any differences among training groups and between women and men in scores on pretests of G*f*. Overall, there were no statistically significant differences between women and men, $F(4, 123) = .64$, $p = .64$, $\eta^2 = .02$, among training groups on pretest scores of G*f*, $F(4, 123) = 1.19$, $p = .27$, $\eta^2 = .04$, or the interaction between sex and training, $F(4, 123) = 1.50$, $p = .10$, $\eta^2 = .05$.

There was also no overall significant sex difference on the G*f* posttest scores, $F(4, 123) = 1.14$, $p = .34$, $\eta^2 = .04$. There were, however, significant differences in posttest scores among training groups, $F(16, 376) = 2.02$, $p = .01$, $\eta^2 = .06$, which may have been a result of training and is analyzed in more detail in the next portion of the Results section. The between-subjects tests indicated there was only one significant difference among training groups for the BETA-III posttest, $F(4, 126) = 5.42$, $p < .001$, $\eta^2 = .15$. Post hoc analyses were conducted to determine if there were any differences between paired training groups. There were a number of significant differences between pairwise training groups for the BETA-III posttest. Overall, the dual *n*-back task outperformed the auditory *n*-back group, $M_{Diff.} = 1.47$, $SE = .60$, $p = .02$; the spatial matrix span group, $M_{Diff.} = 1.14$, $SE = .58$, $p = .05$; and the control group, $M_{Diff.} = 2.28$, $SE = .59$, $p < .001$; the visual *n*-back outperformed the control group, $M_{Diff.} = 1.58$, $SE = .59$, $p = .007$; and the auditory *n*-back group outperformed the control group, $M_{Diff.} = 1.22$, $SE = .61$, $p = .04$. No significant differences existed between pairwise groups for scores on the Cattell posttest and one marginally significant difference existed in the WASI posttest scores — the dual *n*-back training group performed better than the visual *n*-back training group, $M_{Diff.} = 1.24$, $SE = .62$, $p = .05$. The dual *n*-back training group's APM posttest scores were also higher compared to the control group $M_{Diff.} = 3.38$, $SE = 1.19$, $p = .005$. The interaction between sex and training for posttest scores of G*f* was not significant, $F(16, 376) = .97$, $p = .49$, $\eta^2 = .03$.

Performance on the cognitive tests was measured as the number of correct responses (see Table 2 for means and standard deviations). For the Lexical Decision Test, the

Table 1
Means and standard deviations for tests of Gf.

	Pretest			Posttest		
	Men	Women	Total	Men	Women	Total
Raven's						
Dual	14.71 (4.21)	15.29 (3.17)	15.00 (3.67)	17.07 (4.01)	18.00 (4.17)	17.54 (4.04)
Visual	15.57 (5.24)	11.20 (3.95)	13.31 (5.05)	17.36 (4.88)	13.47 (4.10)	15.34 (4.83)
Auditory	14.25 (3.86)	13.54 (3.62)	13.88 (3.68)	15.42 (3.82)	15.23 (4.59)	15.32 (4.15)
STM	13.62 (4.81)	13.07 (3.35)	13.32 (4.02)	15.92 (6.45)	15.13 (4.09)	15.50 (5.22)
Control	13.67 (3.45)	14.43 (3.06)	14.08 (3.20)	14.33 (3.60)	14.00 (3.19)	14.15 (3.32)
Total	14.40 (4.32)	13.46 (3.63)	13.91 (3.99)	16.09 (4.68)	15.14 (4.23)	15.60 (4.50)
Cattell						
Dual	26.64 (4.86)	27.57 (3.67)	27.11 (4.25)	27.93 (5.72)	29.57 (4.33)	28.75 (5.05)
Visual	27.79 (5.13)	24.13 (4.03)	25.90 (4.88)	29.64 (4.77)	25.67 (4.10)	27.59 (4.80)
Auditory	27.17 (5.91)	23.85 (3.95)	25.44 (5.16)	27.75 (4.73)	25.54 (4.37)	26.60 (4.59)
STM	25.77 (4.59)	26.13 (2.42)	25.96 (3.52)	26.38 (5.66)	26.60 (4.40)	26.50 (4.93)
Control	25.33 (4.12)	25.50 (5.02)	25.42 (4.54)	25.42 (3.99)	27.00 (5.08)	26.26 (4.59)
Total	26.57 (4.88)	25.45 (4.02)	25.99 (4.47)	27.49 (5.10)	26.87 (4.57)	27.17 (4.82)
WASI						
Dual	20.50 (2.65)	20.79 (2.75)	20.64 (2.66)	21.71 (2.64)	22.36 (2.02)	22.04 (2.33)
Visual	20.14 (3.28)	18.60 (2.92)	19.34 (3.14)	21.29 (2.67)	20.33 (1.88)	20.79 (2.30)
Auditory	21.33 (1.97)	20.08 (2.06)	20.68 (2.08)	22.17 (2.12)	21.15 (2.34)	21.64 (2.25)
STM	19.38 (2.47)	20.40 (2.26)	19.93 (2.37)	21.08 (2.47)	21.67 (2.58)	21.39 (2.50)
Control	20.67 (2.71)	20.71 (1.64)	20.69 (2.15)	21.33 (2.15)	20.93 (2.20)	21.12 (2.14)
Total	20.38 (2.66)	20.10 (2.46)	20.24 (2.55)	21.51 (2.39)	21.28 (2.26)	21.39 (2.32)
BETA						
Dual	19.50 (2.53)	20.79 (2.12)	20.14 (2.38)	21.86 (2.28)	22.36 (2.33)	22.11 (2.01)
Visual	20.71 (3.20)	19.20 (2.11)	19.93 (2.75)	22.07 (1.90)	20.00 (1.46)	21.00 (1.96)
Auditory	20.58 (1.88)	20.46 (1.66)	20.52 (1.73)	21.00 (2.13)	20.31 (1.93)	20.64 (2.02)
STM	21.15 (1.77)	19.67 (2.13)	20.36 (2.08)	21.54 (2.03)	20.47 (2.59)	20.96 (2.36)
Control	20.00 (1.60)	19.93 (1.98)	19.96 (1.78)	19.58 (2.84)	19.29 (2.20)	19.42 (2.47)
Total	20.38 (2.32)	19.99 (2.04)	20.18 (2.18)	21.26 (2.35)	20.48 (2.22)	20.85 (2.30)

number correct is reported only for the 20 words and does not include nonwords because we were interested in people's ability to access the lexicon quickly and accurately and whether training improves people's ability to access the lexicon. The MANOVA indicated a significant sex difference in the number correct on the cognitive pretests, F (4, 123) = 3.22, $p = .02$, $\eta^2 = .10$, but no significant difference among training groups, F (16, 376) = .39, $p = .98$, $\eta^2 = .01$. Sex differences were present in two of the four cognitive tests. As expected, men outperformed women on the Mental Rotation Test, F (1, 126) = 11.14, $p = .001$, $\eta^2 = .08$. However, there was an unexpected finding that men scored higher on the Extended Range Vocabulary Test than women, F (1, 126) = 6.31, $p = .01$, $\eta^2 = .05$. A significant sex by training interaction was also found, F (16, 376) = 1.82, $p = .03$, $\eta^2 = .06$. The interaction was a result of men in the visual *n*-back group scoring higher than women in the visual *n*-back group on the Paper Folding Test, F (1, 126) = 4.87, $p = .001$, $\eta^2 = .13$; no differences between men and women existed in the other training conditions.

A MANOVA showed a significant sex difference for the cognitive posttests, F (4, 123) = 5.34, $p < .001$, $\eta^2 = .12$, and no significant differences among training groups, F (16, 376) = 1.09, $p = .36$, $\eta^2 = .03$. The interaction between sex and training found in the pretest sessions was not found in the posttest session, F (16, 376) = 1.55, $p = .08$, $\eta^2 = .05$. There was one new significant sex difference present in the cognitive posttests that did not exist in the pretests. In the Paper Folding Pretest, there was an interaction between sex and training, however, there was a main effect of sex found in the Paper Folding Posttest, F (1, 126) = 4.55, $p = .04$, $\eta^2 = .04$ (see Table 2 for means and SDs). Differences in the cognitive posttests were still present in the Mental Rotation Test, F (1, 126) = 20.06, $p = .01$, $\eta^2 = .14$, and the Extended Range Vocabulary Test, F (1, 126) = 6.69, $p = .01$, $\eta^2 = .05$.

Mean RTs for correct answers for each cognitive pretest and posttest were calculated (see Table 3). For the pretest RTs, a MANOVA showed no overall statistical significance between men and women, F (4, 123) = .93, $p = .45$, $\eta^2 = .03$; among training groups, F (16, 376) = .89, $p = .59$, $\eta^2 = .03$; or the interaction between sex and training, F (16, 376) = 1.01, $p = .44$, $\eta^2 = .03$. For the posttest RTs, a MANOVA indicated that there was no overall significant sex difference in RTs for correct answers on the cognitive posttests, F (4, 123) = 2.06, $p = .09$, $\eta^2 = .06$. There were, however, differences among the training groups, F (16, 376) = 1.88, $p = .02$, $\eta^2 = .06$. Post hoc analysis indicated a number of significant pairwise differences among the training groups for posttest RTs in the Extended Range Vocabulary Test, the Lexical Decision Test, and the Paper Folding Test (see Appendix A for details). In general, differences existed between the dual and visual *n*-back training groups for the Extended Range Vocabulary Test with the dual *n*-back training group generally having faster posttest RTs than the visual *n*-back training group. There were no specific patterns in differences for the posttest Lexical Decision Test. Surprisingly, the control group had significantly faster posttest RTs for correct answers on the

Table 2
Means and standard deviations for cognitive tests.

	Pretest			Posttest		
	Men	Women	Total	Men	Women	Total
Vocabulary						
Dual	23.57 (6.21)	17.79 (6.58)	20.68 (6.93)	24.57 (6.17)	19.86 (6.06)	22.21 (6.47)
Visual	20.79 (8.16)	16.00 (7.42)	18.31 (8.02)	21.79 (7.44)	17.00 (7.11)	19.31 (7.55)
Auditory	20.75 (9.12)	17.31 (6.37)	18.96 (7.84)	21.08 (8.82)	18.31 (6.22)	19.64 (7.55)
STM	19.23 (7.84)	20.11 (10.02)	19.71 (8.92)	20.62 (7.49)	21.13 (8.90)	20.89 (8.13)
Control	20.25 (6.94)	16.93 (6.39)	18.46 (6.73)	21.58 (5.99)	17.64 (5.46)	19.46 (5.94)
Total	20.97 (7.60)	17.64 (7.46)	19.23 (7.68)	21.98 (7.15)	18.80 (6.88)	20.32 (7.16)
Lexical						
Dual	19.71 (.47)	19.79 (.43)	19.75 (.44)	19.71 (.47)	19.86 (.36)	19.79 (.42)
Visual	19.71 (.47)	19.73 (.46)	19.72 (.45)	19.86 (.36)	19.87 (.35)	19.86 (.35)
Auditory	19.75 (.45)	19.77 (.44)	19.76 (.44)	19.83 (.39)	19.62 (.51)	19.72 (.46)
STM	19.77 (.44)	19.73 (.46)	19.75 (.44)	19.62 (.51)	19.73 (.46)	19.68 (.48)
Control	19.83 (.39)	19.57 (.51)	19.69 (.47)	19.83 (.39)	19.79 (.43)	19.81 (.40)
Total	19.75 (.43)	19.72 (.45)	19.74 (.44)	19.77 (.42)	19.77 (.42)	19.77 (.42)
Mental rotation						
Dual	15.43 (4.96)	14.21 (6.22)	14.82 (5.55)	16.50 (4.36)	15.21 (5.66)	15.86 (5.00)
Visual	17.71 (4.98)	10.73 (3.90)	14.10 (5.63)	18.29 (4.91)	11.53 (5.97)	14.79 (6.39)
Auditory	16.42 (5.65)	12.23 (5.33)	14.24 (5.78)	19.33 (5.85)	13.38 (5.88)	16.24 (6.50)
STM	13.85 (6.30)	13.27 (5.11)	13.54 (5.59)	16.54 (5.39)	13.87 (5.33)	15.11 (5.43)
Control	14.00 (4.05)	12.50 (3.16)	13.19 (3.60)	16.25 (4.61)	12.71 (4.07)	14.35 (4.60)
Total	15.52 (5.29)	12.58 (4.85)	13.99 (5.26)	17.37 (5.02)	13.32 (5.41)	15.26 (5.59)
Paper folding						
Dual	10.75 (4.01)	13.08 (3.22)	11.92 (3.76)	13.38 (3.88)	14.44 (2.59)	13.91 (3.28)
Visual	14.44 (2.66)	8.98 4.04)	11.62 (4.37)	15.41 (2.05)	11.44 (3.60)	13.36 (3.54)
Auditory	13.51 (2.65)	11.26 (3.50)	12.34 (3.27)	15.09 (2.26)	11.01 (4.09)	12.97 (3.88)
STM	10.36 (4.82)	11.30 (3.34)	10.86 (4.04)	11.82 (3.70)	12.49 (4.05)	12.18 (3.84)
Control	12.11 (3.87)	11.13 (3.66)	11.58 (3.72)	11.58 (4.09)	11.58 (3.35)	11.58 (3.63)
Total	12.23 (3.92)	11.12 (3.71)	11.65 (3.84)	13.49 (3.57)	12.20 (3.68)	12.82 (3.68)

Table 3
Means and standard deviations for cognitive tests RTs.

	Pretest			Posttest		
	Men	Women	Total	Men	Women	Total
Vocabulary						
Dual	7860 (2360)	8251 (2665)	8055 (2478)	7001 (2138)	6774 (2691)	6888 (2388)
Visual	10324 (3887)	10439 (3934)	10384 (3846)	7886 (2508)	9207 (3789)	8569 (3248)
Auditory	8845 (2009)	10900 (4993)	9913 (3926)	7607 (2677)	9204 (4349)	8437 (3661)
STM	9056 (2735)	8991 (3583)	9021 (3159)	7645 (2399)	6659 (2318)	7117 (2365)
Control	9143 (2618)	8596 (3133)	8848 (2863)	6552 (1872)	6642 (1851)	6601 (1824)
Total	9049 (2854)	9422 (3764)	9244 (3353)	7349 (2315)	7683 (3272)	7523 (2849)
Lexical						
Dual	821 (125)	768 (153)	795 (140)	763 (135)	725 (131)	744 (132)
Visual	818 (121)	806 (117)	812 (117)	811 (119)	823 (127)	818 (121)
Auditory	803 (109)	782 (117)	792 (111)	821 (136)	721 (79)	769 (119)
STM	813 (132)	800 (107)	806 (117)	724 (90)	731 (80)	728 (83)
Control	803 (112)	780 (111)	791 (110)	787 (131)	724 (56)	753 (101)
Total	812 (117)	788 (119)	799 (118)	781 (124)	746 (105)	763 (115)
Mental Rotation						
Dual	19185 (7921)	17192 (8884)	18189 (8321)	15180 (7472)	13481 (7071)	14331 (7190)
Visual	18455 (6157)	14005 (6939)	16153 (6841)	15892 (8798)	9760 (6483)	12721 (8166)
Auditory	17561 (7954)	16688 (7472)	17107 (7557)	14267 (6687)	15741 (12264)	15033 (9811)
STM	12613 (7329)	17364 (8358)	15158 (8119)	13599 (6255)	15086 (9081)	14396 (7792)
Control	16202 (6317)	13865 (5912)	14959 (6091)	13265 (4099)	11737 (7257)	12442 (5948)
Total	16863 (7337)	15812 (7536)	16314 (7433)	14495 (6779)	13104 (8642)	13769 (7810)
Paper folding						
Dual	18256 (8843)	17368 (4276)	17862 (6834)	18755 (7219)	13904 (4098)	16330 (6267)
Visual	20060 (5840)	14881 (5044)	17381 (5958)	17137 (6393)	14138 (7330)	15586 (6940)
Auditory	16360 (5137)	19885 (5726)	18193 (5632)	15367 (8166)	15138 (5457)	15248 (6743)
STM	15260 (6842)	16944 (5517)	16162 (6109)	15559 (5316)	14988 (7033)	15253 (6188)
Control	17756 (6320)	14166 (4823)	15823 (5745)	11264 (3096)	12426 (4303)	11890 (3767)
Total	17624 (6771)	16583 (5336)	17081 (6064)	15759 (6595)	14117 (5765)	14902 (6207)

Paper Folding Test than any of the training groups, which no explanation can be provided except that the improvements were due to chance. No pairwise differences were present in the posttest Mental Rotation Test. No significant interaction between sex and training was present for RTs on the cognitive posttests, $F\,(16, 376) = 1.27$, $p = .22$, $\eta^2 = .04$.

3.1. Training performance

Training performance was analyzed for any potential differences between men and women's training performance or differences among groups. Because Jaeggi et al. found no difference between participants who completed 17 vs. 19 sessions and there are missing data in the current study for sessions 19 and 20, only sessions 1 through 18 were analyzed. Average performance for the *n*-back groups is defined as the average number of *n*-back a person reached for that training session. For the spatial matrix training group, average performance is defined as the average number of blue circles in a sequence remembered for that training session.

There was a significant change in daily average performance over the 18 training sessions for all training groups, $F\,(17, 86) = 18.42$, $p < .001$, $\eta^2 = .79$. There was no significant two-way interaction between sex and change in the daily average performance over the 18 sessions, $F\,(17, 86) = 1.21$, $p = .28$, $\eta^2 = .19$. However, there was a significant two-way interaction between training groups and change in the daily average performance, $F\,(51, 257) = 1.56$, $p = .009$, $\eta^2 = .24$. To determine *where* the differences existed among training groups, the daily average performance for training session 1 was subtracted from training session 18 to calculate the total improvement made in each group. Between-subjects LSD unadjusted post-hoc analyses were conducted to determine differences among training groups. The post-hoc analyses revealed that no statistically significant differences existed among the three *n*-back training groups, but that differences existed between each *n*-back training group and the spatial matrix span group. In general, the participants in the spatial matrix span group had a higher average performance compared to the three *n*-back training groups. The potential 3-way interaction that includes training sessions, sex, and training groups was not significant, $F\,(51, 257) = .99$, $p = .51$, $\eta^2 = .16$.

Overall, participants in each training group experienced improvements, but the improvements were not as strong compared to Jaeggi et al. (2008). Participants were not paid in the current study and could have had an impact on their motivation. Other external motivators, aside from monetary motivation, may have been needed as well. Although the experimenters provided verbal motivation every session, it may not have been enough to keep people motivated and may have resulted in the lower average *n*-back performances.

3.2. Gains in Gf and cognitive abilities

3.2.1. Tests of Gf

To make an accurate comparison to Jaeggi et al. (2008, 2010), paired *t*-tests comparing pretest and posttest scores were conducted only for the participants in the dual *n*-back training group. Participants who completed the dual *n*-back training program experienced improved scores on the APM, $t\,(27) = 6.19$, $p < .001$, $d = 1.17$, CI [1.69, 3.38]; Cattell's test, $t\,(27) = 2.99$, $p = .006$, $d = .56$, CI [.56, 2.77]; the WASI subtest, $t\,(27) = 3.98$, $p < .001$, $d = 1.00$, CI [.67, 2.11]; and the BETA-III subtest, $t\,(27) = 4.18$, $p < .001$, $d = .79$, CI [1.00, 2.98]. Based on these analyses, Jaeggi et al.'s results were replicated (i.e., Jaeggi et al. found an effect size of .65) to the extent that there was improvement in scores on the APM.

A repeated measures analysis of the scores on tests of G*f* revealed a significant difference between pretest scores and posttest scores, $F\,(4, 123) = 28.73$, $p < .001$, $\eta^2 = .48$: posttest scores were generally higher than pretest scores. There was no difference between men and women's pretest and posttest scores, $F\,(4, 123) = .57$, $p = .69$, $\eta^2 = .02$. Training, however, did have a significant effect on the differences between pretest and posttest scores, $F\,(16, 504) = 2.26$, $p = .004$, $\eta^2 = .07$. There was no interaction effect of sex and training on the difference in scores on pretests and posttests, $F\,(16, 504) = .46$, $p = .96$, $\eta^2 = .01$. Based on these findings, the hypothesis that women would experience greater gains on tests of G*f* as a result of training was not supported. Furthermore, there is no evidence to support the hypothesis that women would experience greater gains if they were in the dual *n*-back or visual *n*-back training groups. Because of the nonsignificant main effect for sex or interaction between sex and training group, no further analyses were conducted to test for sex differences in gains on tests of G*f*.

Results of the repeated measures analysis indicated that training had a significant impact on the difference in scores between pretests and posttests. Surprisingly, type of training had a significant effect on the difference in pretest and posttest scores (see Table 1 for means and SDs) for only the APM, $F\,(4, 126) = 3.49$, $p = .01$, $\eta^2 = .10$, and the BETA-III pretest and posttest scores, $F\,(4, 126) = 4.86$, $p = .001$, $\eta^2 = .13$. Training type did not have a statistically significant effect on the difference for pretest and posttest scores on the Cattell test, $F\,(4, 126) = .91$, $p = .46$, $\eta^2 = .03$, or the WASI, $F\,(4, 126) = 1.33$, $p = .26$, $\eta^2 = .04$. Even though there were no overall statistically significant differences between pretest and posttest scores on the Cattell and WASI tests, contrasts were conducted on Cattell and WASI first to determine if there were any specific differences among training groups. Post hoc analyses were not used because there were a priori hypotheses for all tests of G*f*, and no adjustments have been made to the computed *p*-values. Contrasts for the WASI and Cattell tests revealed no differences among any of the training groups.

Although training types may not have differed from one another in the improvement of scores on the Cattell and WASI tests, there was generally significant improvement as shown by the one-sample *t*-tests to compare the mean gains experienced by each training group per test of G*f* to the null hypothesis of zero gain (see Appendix B). Based on the current statistical evidence, Jaeggi et al.'s (2008, 2010) studies were, in general, corroborated. However, there are qualifications that need to be made for the claim that training on the dual *n*-back task improves G*f*. The current evidences suggests that specific types of training have different effects on the improvement in scores on the APM and BETA-III tests; thus, specific a priori contrasts were conducted only on the APM and BETA-III to test predicted group differences.

Based on the a priori contrasts for the APM test, the hypothesis that the participants who trained using the dual *n*-back task experienced greater gains than the control group ($M_{Diff} = 2.46$, $SEM = .70$) was supported, $t\ (131) = 3.49$, $p < .001$, $d = .61$. Also, in support of our hypotheses, the visual *n*-back task group experienced greater gains than the control group ($M_{Diff} = 1.96$, $SEM = .70$) was supported, $t\ (131) = 2.80$, $p = .006$, $d = .49$. However, the current evidence shows that participants do not have to train using a WMC task as training because, surprisingly, participants in the spatial matrix span training condition also experienced a significant improvement on the APM compared to the control group ($M_{Diff} = 2.10$, $SEM = .70$), $t\ (131) = 2.98$, $p = .003$, $d = .52$. The evidence that participants in the auditory *n*-back training condition did not differ in gains compared to the control group, but that the other *n*-back training groups experienced greater gains than the control group partially supports the hypothesis that in order to experience gains in G*f*, participants need to train using a program that contains a visuospatial component. Furthermore, in a serendipitous and quite intriguing finding, participants in the spatial matrix span training also experienced greater gains on the APM than the control group suggesting that WMC training task may not be necessary, but a visuospatial component may be. However, there were no differences in gains on the APM among training groups, which suggests that completing any type of training may be better than doing nothing.

The repeated measures analysis also showed an overall improvement in the BETA-III subtest and the same contrasts used for the other tests of G*f* were conducted. The dual *n*-back training group experienced greater improvements ($M_{Diff} = 2.50$, $SEM = .60$) than the control group, $t\ (131) = 4.18$, $p < .001$, $d = .73$. Participants in the visual *n*-back group also experienced greater gains ($M_{Diff} = 1.61$, $SEM = .59$) than the control group, $t\ (131) = 2.70$, $p = .008$, $d = .47$. However, the auditory *n*-back training group and the spatial matrix span training group did not differ from the control group in the improvements. Furthermore, the auditory *n*-back training group ($M_{Diff} = 1.84$, $SEM = .61$) experienced significantly less improvement than the dual *n*-back training group, $t\ (131) = 3.05$, $p < .001$, $d = .53$. The spatial matrix span training group also experienced significantly less improvement ($M_{Diff} = 1.36$, $SEM = .59$) than the dual *n*-back group, $t\ (131) = 2.31$, $p = .02$, $d = .40$. There were no differences between the auditory *n*-back group or the spatial matrix group and the visual *n*-back training group. The evidence from the contrasts for the BETA-III subtest suggests that the dual *n*-back training was superior for making gains followed by the visual *n*-back training. The auditory *n*-back training did little in improving scores on the BETA-III subtest and the spatial matrix span may or may not improve scores on the BETA-III subtest.

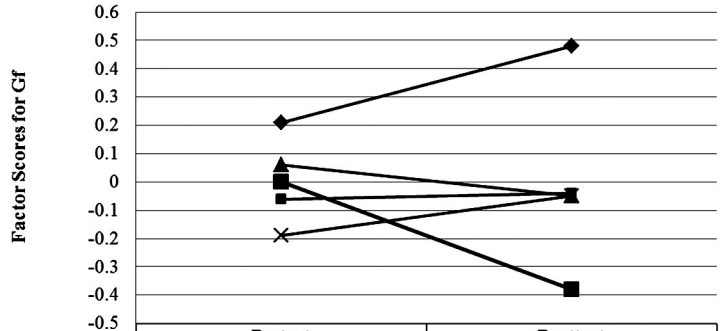
So far, the results have indicated that training using the dual and visual *n*-back task and the spatial matrix training improves scores on two of the four tests independently, but does not indicate whether there was a change in the latent factor G*f*. Thus, the four tests of G*f* were combined into a latent factor for the pretest and posttest versions. The pretest latent factor of G*f* was a good fit, $\chi^2_{(6)} = 223.09$, CFI = .99, TLI = .98, RMSEA = .070; as was the posttest latent factor of G*f*, $\chi^2_{(6)} = 210.57$, CFI = 1.00, TLI = 1.03, RMSEA = .000. However, when the two latent factors were combined into a non-restricted model, the model did not remain stable, $\chi^2_{(6)} = 720.24$, CFI = .90, TLI = .85, RMSEA = .164. Even when the two latent factors were combined into a fixed model, the model remained unstable, $\chi^2_{(6)} = 720.24$, CFI = .90, TLI = .88, RMSEA = .149. The instability of the model containing the two latent factors may suggest a change between the pretest latent factor and the posttest latent factor. However, the instability in the model may also be a result of simply creating a latent factor from two different testing periods.

Standardized scores from the factor analyses were saved for the pretest and posttest latent factors. The posttest standardized scores were subtracted from the pretest standardized scores to create a difference score and were compared using a one-way ANOVA to test for differences among training groups (see Fig. 3). There was a significant difference among the training groups difference in pretest and posttest standardized scores, $F\ (4, 131) = 6.56$, $p < .001$, $\eta^2 = .17$. Post hoc tests showed that the dual *n*-back training group experienced greater gains in their latent factor scores than the auditory training group, $t\ (131) = -2.69$, $p = .008$, $d = .47$; and greater gains over the control group, $t\ (131) = -4.74$, $p < .001$, $d = .83$. The visual *n*-back training group also experienced greater gains in their latent factor scores compared to the control group, $t\ (131) = -3.86$, $p < .001$, $d = .67$. The final pairwise comparison that differed significantly was the spatial matrix training group experiencing greater gains than the control group, $t\ (131) = -2.80$, $p = .005$, $d = .49$.

Overall, the pairwise differences between training groups for latent factor standardized scores were similar in terms of which training groups experienced greater gains. The dual and visual *n*-back training groups and the spatial matrix training group experienced greater gains than the control group whereas the auditory *n*-back training group did not. However, there was no difference between the auditory *n*-back training and the visual *n*-training group or the spatial matrix training group, but did differ with the dual *n*-back training group. Therefore, training with the dual or visual *n*-back tasks or the spatial matrix task is better than doing nothing, but may not be better than the auditory *n-back* training with the exception of the dual *n*-back training.

3.2.2. Cognitive tests (accuracy)

According to a repeated measures analysis, the between-subjects analyses showed a sex difference in the number of correct answers on the four cognitive tests, $F\ (4, 123) = 4.81$, $p < .001$, $\eta^2 = .14$, no difference among training groups, $F\ (16, 504) = .76$, $p = .74$, $\eta^2 = .02$, and there was a marginal significant interaction between sex and training, $F\ (16, 376) = 1.62$, $p = .06$, $\eta^2 = .05$. Because there was a sex difference, an examination of the tests of between-subjects effects was necessary. Sex differences existed for the Extended Range Vocabulary Test, $F\ (1, 126) = 6.88$, $p = .01$, $\eta^2 = .05$, the Mental Rotation Test, $F\ (1, 126) = 17.15$, $p < .001$, $\eta^2 = .12$, and the Paper Folding Test, $F\ (1, 126) = 4.37$, $p = .04$, $\eta^2 = .03$. Men had a higher number of correct answers for the Extended Range Vocabulary Test, the Mental Rotation Test, and the Paper Folding Test than women (see Table 2 for means and SDs). There was no sex difference for the Lexical Decision Test, $F\ (1, 126) = .10$, $p = .76$, $\eta^2 = .00$.



	Pretest	Posttest
Dual N-Back	0.21 (.95)	0.48 (.98)
Visual N-Back	-0.19 (1.26)	-0.05 (1.00)
Auditory N-Back	0.06 (.96)	-0.05 (.89)
Spatial Matrix	-0.06 (.94)	-0.04 (1.09)
Control	0 (.84)	-0.38 (.88)

Fig. 3. Change in pretest and posttest latent factor scores for each training group. Factor scores were saved as regression weights.

Although the between-subjects interaction between sex and training was not significant, the marginally significant value justified a closer look at each test to determine if there was an interaction in one of the cognitive tests. There was an interaction between sex and training for the Paper Folding Test, $F\ (4,\ 126) = 4.75$, $p = .001$, $\eta^2 = .13$. Women in the dual *n*-back training group and the spatial matrix span training group had a higher number of correct answers than men in those groups, whereas men in the visual and auditory *n*-back training groups had a higher number of correct answers than women in the same groups (see Table 2 for means and SDs).

A MANOVA showed a significant difference between the cognitive pretests and posttests, $F\,(4, 123) = 13.18$, $p < .001$, $\eta^2 = .30$. An analysis of the univariate tests indicated a significant change in performance on the Extended Range Vocabulary Test, $F\,(1, 126) = 12.56$, $p = .001$, $\eta^2 = .09$; the Mental Rotation Test, $F\,(1, 126) = 21.52$, $p < .001$, $\eta^2 = .15$; and the Paper Folding Test, $F\ (1,\ 126) = 23.30$, $p < .001$, $\eta^2 = .16$. There was no significant difference in the number of correct answers between the pretest and posttest for the Lexical Decision Test, $F\ (1,\ 126) = .50$, $p = .48$, $\eta^2 = .00$. Overall, the three tests in which there were significant changes showed more correct answers on the posttest than the pretest.

There was no effect of sex on gains in number of correct answers on cognitive tests with the exception of the Mental Rotation Test, $F\,(1, 126) = 4.02$, $p = .047$, $\eta^2 = .03$. Surprisingly, the sex difference was not a result of women improving their performance more than men; instead, men ($M_{Gain} = 1.85$, $SEM = .41$) improved significantly more than women ($M_{Gain} = .75$, $SEM = .40$). Change in number of correct answers was marginally significant due to training only for the Paper Folding Test, $F\,(4, 126) = 2.42$, $p = .051$, $\eta^2 = .07$; training group effects were not significant for changes in performance on the other tests. A final test of the potential three-way interaction between changes in percentage correct, sex, and training did not exist for any of the cognitive tests, $F\,(16, 504) = .86$, $p = .49$, $\eta^2 = .03$.

Overall, participants in the dual *n*-back training condition experienced greater improvements compared to the control group. However, the improvements made by participants in the dual *n*-back training group did not differ significantly from the other training groups. Thus, it could be inferred that the dual *n*-back training is better than doing nothing, but that the other three training tasks may provide similar results. Models for the latent factors of verbal ability (i.e., Extended Range Vocabulary Test and Lexical Decision Test) and visuospatial ability (i.e., Mental Rotation Test and Paper Folding Test) were not stable. The models were not stable partially because there were only two measures per factor; thus, further analyses were not conducted on the latent factors.

Based on the evidence presented thus far, the hypothesis that there would be improvements on the Mental Rotation Test and Paper Folding Test and not on the Extended Range Vocabulary Test and Lexical Decision Test was partially supported. The hypothesis was correct in that there were significant improvements on the Mental Rotation Test and Paper Folding Test. However, the hypothesis was not fully supported because there were also significant improvements in the Extended Range Vocabulary Test. As predicted, there was no significant improvement on the Lexical Decision Test. The hypothesis also stated that the improvements in the tests would vary with types of training; however, this was not the case as there was not statistically significant evidence that type of training had an effect on the changes between the pretests and posttests with the exception of the Paper Folding Test.

The hypothesis stating that improvements on the Mental Rotation Test and Paper Folding Test would be the result of the dual or visual *n*-back training was also partially supported in that participants in the dual *n*-back training condition improved only on the Paper Folding Test, but this could be an isolated finding. Finally, the hypothesis stating that women would experience greater improvement in visuospatial abilities than men was not supported by the current statistical evidence which, in fact, went in the opposite direction such that men

experienced greater improvements on the Mental Rotation Test than the women.

3.2.3. Cognitive tests (RTs)

Training was hypothesized to improve RTs on the cognitive tests. A repeated measure analysis showed no between-subjects effect for sex, $F\ (4, 123) = 1.99$, $p = .10$, $\eta^2 = .06$, or training, $F\ (16, 376) = 1.32$, $p = .18$, $\eta^2 = .04$. The potential sex by training interaction was also nonsignificant, $F\ (16, 376) = .96$, $p = .50$, $\eta^2 = .03$. The analysis for within-subjects effect did show that there was a difference between pretests and posttests RTs, $F\ (4, 123) = 20.48$, $p < .001$, $\eta^2 = .40$, with the posttests having faster RTs. However, there was no sex difference in the change in RTs, $F\ (4, 123) = .14$, $p = .97$, $\eta^2 = .00$, or differences among training groups, $F\ (16, 376) = 1.11$, $p = .35$, $\eta^2 = .04$. The potential 3-way interaction between change in RTs, sex, and training was also nonsignificant, $F\ (16, 376) = 1.19$, $p = .28$, $\eta^2 = .04$. Because of the change in RTs between cognitive pretests and posttests, an inspection of the univariate tests was necessary. The univariate tests indicated faster posttest RTs (see Table 3 for means and SDs) for the Extended Range Vocabulary Test, $F\ (1, 126) = 56.23$, $p < .001$, $\eta^2 = .31$; the Lexical Decision Test, $F\ (1, 126) = 10.65$, $p < .001$, $\eta^2 = .08$; the Mental Rotation Test, $F\ (1, 126) = 21.55$, $p < .001$, $\eta^2 = .15$; and the Paper Folding Test, $F\ (1, 126) = 15.82$, $p < .001$, $\eta^2 = .11$. Overall, there were no sex differences and no differences among training groups for improving RTs for the cognitive tests and the observed improvements were most likely due to practice effects.

4. Discussion

The primary goal of our study was to test the hypothesis that scores on tests of G*f* would improve only for participants who had a visuospatial component in training to improve WMC. Overall, we found this hypothesis to be supported, but with a surprising finding that a visuospatial STM training program was also beneficial. We also tested for differences between men and women's gains on their scores on tests of G*f* with the hypothesis that women would experience greater gains, which did not turn out to be case. Our study corroborates Jaeggi et al.'s (2008, 2010) findings that the dual *n*-back task is a viable training program for improving scores on select tests of G*f*. However, the results of the current study go beyond our hypotheses and raise many questions about how researchers define G*f* regarding plasticity in terms of improvement through training and the psychometric tests that are used to test G*f*.

G*f* is conceptually defined as an ability to solve novel problems without the use of prior strategies, and G*f* is assumed to function without relying on strategies derived from verbal and visuospatial abilities (Horn & Cattell, 1966; Raven, 2000). However, recent research on tests of G*f* has provided evidence that the tests may not be testing G*f* exclusively (Schweizer et al., 2007) and that there is a strong relationship between visuospatial abilities and G*f*. The current study lends support for multidimensionality in tests of G*f* based on the evidence that the greatest gains in the tests of G*f* were experienced by training groups that had a visuospatial component including the WMC *and* the STM training. Furthermore, the Paper Folding Test was the only cognitive test in which participants experienced gains as a result of the dual *n*-back training. These findings raise an important question: Did improvements on the Paper Folding Test result from an improvement in G*f*, an improvement in visuospatial skills, or are these merely the same labels for the same construct? Tests of G*f* are assumed to predict other cognitive abilities (Sternberg, 2008). If there was an improvement in G*f*, then there should have also been an improvement in the Mental Rotation Test because G*f* is not domain or task specific.

One of the primary concerns of the current study is whether G*f* was improved or if the improvement in scores on tests of G*f* was really an improvement of test taking abilities for those specific types of tests. It could be that the training simply allowed participants to improve their test taking abilities just enough to show a significant improvement over practice effects. In either case, as observed in the current study, a visuospatial component is needed in the training program for participants to experience those improvements.

4.1. Improving scores on tests of G*f*

The question, "Does training to enhance working memory capacity improve scores on the APM," can be answered with a "yes." However, the improvements in scores on other tests of G*f* are limited to the BETA-III subtest. The improvement in scores on the APM and BETA-III subtest was the result of participants completing training programs to improve WMC that contained a visuospatial component; as seen by the participants who completed the dual and visual *n*-back training. However, participants who completed the STM training also experienced greater gains than the control group and did not differ in gains compared to the WMC training groups. Why did participants in the STM training group experience gains similar to the WMC training groups? One explanation could be that the STM training enhanced a visuospatial mechanism shared by STM and WM. However, we did not have an auditory STM task to substantiate this explanation. Thus, the more likely explanation that STM training had an effect is because the STM training enhanced the shared short-term storage component that influences G*f* as shown by Colom et al. (2006), Martínez et al. (2011), Krumm et al. (2009), and Hornung et al. (2011).

The constructs STM, WMC, executive functioning, attention, and G*f* *do* have a common factor: short-term capacity (Halford, Cowan, & Andrews, 2007). Perhaps, what the cognitive training is truly doing is expanding participants' limited capacity that all of the constructs have in common, but this still does not explain why training with a visuospatial component is necessary. One possibility is that because visuospatial processing is more complex than verbal processing, that training visuospatial abilities is having a unique effect on people's test taking abilities on tests of G*f*.[2] Furthermore, does the complexity of visuospatial skills mean, however, that improving visuospatial skills abilities transfer to a *verbal* test of G*f*? Of course, without empirical evidence, a solid conclusion cannot be made, but we would speculate that there would be a

[2] We would like to thank an anonymous reviewer for bringing this idea to our attention.

transfer of ability for two reasons. First, studies by Colom et al. (2006) showed that regardless of the types of STM and WM measures, they all shared the common component of storage capacity. If visuospatial training increases storage it is possible the storage is general for different types of information. Second, training with a visuospatial component may help people encode verbal information visually and result in a better storage capacity, which would lead to higher scores. Finally, would verbal STM training have the same effect as the visuospatial STM training? Based on the evidence from the auditory *n*-back training group, we would suspect that verbal STM training would not improve scores on tests of G*f*. Perhaps, verbal STM training would improve scores on a verbal G*f* test and should be considered for future research.

Although training STM or WMC improves scores on the APM, the training that contains a visuospatial component should experience the greatest gains. Participants who completed the auditory *n*-back training experienced marginal gains compared to the control group, but did not differ from the other training groups. There were also differences among training groups depending on the test of G*f*. The current study suggests that the dual and visual *n*-back training programs are superior to the auditory *n*-back and spatial matrix span training programs because the dual and visual *n*-back tasks are more complex tasks that require greater effort, especially the dual *n*-back task, *and* they have a visuospatial component. The third "best" training program is the STM training. Although the STM training is not as complicated as n-back training, it still requires effort in a visuospatial task that may have led to improvements in scores on tests of G*f*, but needs to be compared to an auditory STM training test in future research. Finally, the auditory *n*-back training provided limited benefits; it definitely does not provide the training necessary for participants to experience strong gains.

The only difference between the auditory *n*-back training and the other WMC training programs is the type of stimuli used to train participants. Because the auditory *n*-back training lacks a visuospatial component, participants who completed this type of training were similar to the control group in that neither group had any training with visuospatial stimuli or experienced significant improvement in scores on tests of G*f*. A visuospatial component most likely provides an advantage in obtaining gains in scores on tests of G*f* that are visuospatial. However, there is also the possibility that because there were no improvements in the auditory *n*-back training group, but there was an improvement in the STM visuospatial training group, that the improvements in scores on tests of G*f* is a result of improving the capacity of the specific modality (i.e., the visuospatial component).

4.2. Conceptualizing and measuring Gf

A primary issue regarding the psychometric tests of G*f* is whether the tests are measuring the same construct. Although the tests of G*f* were all significantly correlated with each other in the current study, there was improvement in only two of the four tests. If the tests of G*f* are truly measuring the same construct, then there should have been a consistent pattern of improvement across all tests of G*f* as a result of training. Even though the APM is correlated with Cattell's Culture Fair Test ($r = .69$ in current study), no gains on the Cattell's were made as a result of specific training. This could have been because Cattell's Culture Fair Test is supposed to, at least psychometrically, test *g* as defined by Spearman (1904) and not G*f* exclusively. The directions given to test takers for Cattell's Culture Fair Test are also different than the APM in that the directions for Cattell's explicitly states the rules for each subtest whereas the directions for the APM do not identify the rules for each problem; thus, the test taker ends up needing to identify the rules in addition to producing an answer. Thus, the APM is a more complex task than what is required by the Cattell Culture Fair Test.

Correlations were also moderate between the APM and the WASI subtest ($r = .61$ in the current study) and, yet, there was only a marginal impact on gains in scores for WASI subtest. A possible explanation for the marginal impact on the WASI subtest scores is that the subtest was only a small portion of a larger test of intelligence (*g*), and may not have strong validity compared to the APM for measuring G*f* exclusively. Although it seems probable that the lack of gains on the WASI subtest was a result of this test being only a part of a larger test and it may not be as valid as a standalone test, this reasoning does not coincide with the impact that training had on the BETA-III subtest because it is also part of a larger test of intelligence, but there were improvements on the BETA-III subtest.

Why would there be an improvement on two tests of G*f* that had a lower correlation value than some of the other tests? When the Beta-III and the WASI were created, they were correlated with other tests using an aggregate score from all of the subtests (Kellogg & Morton, 1999; WASI Manual, 1999). In the current study, only one subtest was used and could have impacted the correlation between the BETA-III subtest and WASI subtest and the APM. In other words, the correlation value could have been affected by having a value from only a portion of the WASI or BETA-III tests instead of the entire test, which could have resulted in a less valid correlation value between the subtests (i.e., WASI and BETA-III) and the tests that were given in their entirety (i.e., APM's and Cattell's). The other explanation is that the training is specific to improving G*f* exclusively as it is measured by the APM, but not other tests or subtests of G*f*. If the training is specific to one test of G*f*, then it is apparent that the tests of G*f* are not measuring the same construct or the training is improving the test taking abilities for a very specific type of test.

4.3. Improving cognitive abilities

The current study provided evidence that there were limited improvements on the two verbal tests and the two visuospatial tests as a result of cognitive training. The Paper Folding Test was the only cognitive test for which participants experienced greater improvements as a result of training when compared to the control group. If scores on the Mental Rotation Test were also improved as a result of training, then a conclusion could have been made that training improves visuospatial abilities. However, there was also an unpredicted improvement in the Extended Range Vocabulary Test that was not a result of training. The improvement in the number of correct on the vocabulary test is most likely a practice effect. Furthermore, training did not have an effect on RTs for any of

the cognitive tests, which could be a result of the training not necessarily having an effect on processing speed. If training does not help with processing speed as observed in the current study, then training probably has an effect on the allocation of resources to the appropriate information while ignoring irrelevant information, which could explain the increase in number of correct answers in the Paper Folding Test. If resources are directed more efficiently, then this would lead to more correct answers, but not necessarily more questions being answered.

4.4. Limitations and future research

A few notable limitations in our study should be empirically tested in future research. Time restrictions on the APM and the WASI subtest are not typically used. If a time restriction is used on the APM, then it is typically set at a 45 minute limitation (Raven et al., 1998). Moody (2009) pointed out that having participants complete items in the APM within a restrictive time limit does not allow them to attempt the more difficult items in the test, which are located toward the end of the test. Furthermore, Moody argued that if the participants do not attempt the difficult items, then the potential high score a participant may receive on the test is less predictable.

Another limitation of the study is the use of the control group to help determine how much of the improvements were due to practice effects. Although there were no improvements experienced by the control group, the training groups could have experienced improved scores as a result of the training improving participants' memory for test items. If participants' memory were improved for remembering test items, then the improvements are not a result of improving G*f*, but improving the practice effect through improving memory.

In a similar vein to using time restrictions on tests of G*f*, the current study did not test if the training has an effect on a larger test of intelligence such as the Stanford–Binet Intelligence Scales, the complete Wechsler Scale of Intelligence, the Woodcock–Johnson III, or the complete BETA-III test. The larger scale tests of intelligence do not test G*f* exclusively, but the improvement in G*f* could lead to solving problems on the larger tests of intelligence with greater ease.

Another limitation of the current study that opens the doors for future research is that there has been no determination for whether the training transfers to everyday situations or habits such as taking notes, improving study habits, improving skills learning, improving test scores in college courses, or improved attention during class lectures. If the training is truly improving attentional control and G*f*, then a reasonable hypothesis would be that a person would be able to take notes, study for tests, and attend to information more efficiently. Tests of G*f* also predict job performance, and the improvement could transfer to those aspects as well. The training may transfer to academic performance or skills performance because exercising cognitive abilities may follow the old adage that, "Doing something is better than doing nothing."

5. Conclusion

The results of the current study provide evidence that G*f* as measured by the APM is more malleable than was previously thought. We were successful in corroborating Jaeggi et al.'s (2008, 2010) studies in that participants' scores improved on the APM and the BETA-III subtest after completing the dual *n*-back and visual *n*-back training programs. However, participants who completed the spatial matrix span training program also experienced gains in the tests of G*f*. The current study, in conjunction with Jaeggi et al.'s studies, has clearly demonstrated that scores on tests of G*f* can be improved after training to improve WMC and can be used as a foundation for future research investigating ways to improve cognitive abilities.

Appendices A and B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.intell.2013.05.006.

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. http://dx.doi.org/10.1037/0033-2909.131.1.30.

Buschkuehl, M., Jaeggi, S. M., Kobel, A., & Perrig, W. J. (2008). *BrainTwister: A collection of cognitive training tasks.* University of Bern, Switzerland: Department of Psychology.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, UK: Cambridge University Press.

Castel, A. D., Pratt, J., & Drummond, E. (2005). The effects of action video game experience on the time course of inhibition of return and efficiency of visual search. *Acta Psychologica*, *119*, 217–230. http://dx.doi.org/10.1016/j.actpsy.2005.02.004.

Cherney, I. D. (2008). Mom, let me play more computer games: They improve my mental rotation skills. *Sex Roles*, *59*, 776–786. http://dx.doi.org/10.1007/s11199-008-9498-z.

Colom, R., Abad, F. J., Rebollo, I., & Shih, P. C. (2005). Memory span and general intelligence: A latent-variable approach. *Intelligence*, *33*, 623–642. http://dx.doi.org/10.1016/j.intell.2005.05.006.

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition*, *34*, 158–171. http://dx.doi.org/10.3758/BF03193395.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505. http://dx.doi.org/10.1080/14640748108400805.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*, 51–57. http://dx.doi.org/10.1177/0963721409359277.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466. http://dx.doi.org/10.1016/S0022-5371(80)90312-6.

Ekstrom, R. B., French, J., Harman, H. H., & Dermen, D. (1976). *Kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology. General*, *128*, 309–331. http://dx.doi.org/10.1037/0096-3445.128.3.309.

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, *18*, 850–855. http://dx.doi.org/10.1111/j.1467-9280.2007.01990.x.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*, 534–537. http://dx.doi.org/10.1038/nature01647.

Green, C. S., & Bavelier, D. (2006a). Effect of action video games on the spatial distribution of visuospatial attention. *Journal of Experimental Psychology. Human Perception and Performance*, *32*, 1465–1478. http://dx.doi.org/10.1037/0096-1523.32.6.1465.

Green, C. S., & Bavelier, D. (2006b). Enumeration versus multiple object tracking: The case of action video game players. *Cognition*, *101*, 217–245. http://dx.doi.org/10.1016/j.cognition.2005.10.004.

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science*, *18*, 88–94. http://dx.doi.org/10.1111/j.1467-9280.2007.01853.x.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences*, *11*, 236–242. http://dx.doi.org/10.1016/j.tics.2007.04.001.

Halpern, D. F. (2012). *Sex differences in cognitive abilities* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Halpern, D. F., & Wai, J. (2007). The world of competitive scrabble: Novice and expert differences in visuospatial and verbal abilities. *Journal of Experimental Psychology. Applied*, *13*, 79–94. http://dx.doi.org/10.1037/1076-898X.13.2.79.

Heitz, R. P., Unsworth, N., & Engle, R. W. (2004). Working memory capacity, attention control, and fluid intelligence. In O. Wilhelm, & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 61–77). Thousand Oaks, CA: Sage Publications, Inc.

Horn, W. (1983). *Leistungsprüfsystem [Performance-Test-System]* (2nd ed.) Göttingen: Hogrefe.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*, 258–270. http://dx.doi.org/10.1037/h0023816.

Hornung, C., Brunner, M., Reuter, R. A. P., & Martin, R. (2011). Children's working memory: Its structure and relationship to fluid intelligence. *Intelligence*, *39*, 210–221. http://dx.doi.org/10.1016/j.intell.2011.03.002.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*, 581–592. http://dx.doi.org/10.1037/0003-066X.60.6.581.

Jaeggi, S. M., Buschkuehl, M., Etienne, A., Ozdoba, C., Perrig, W. J., & Nirkko, A. C. (2007). On how high performers keep cool brains in situations of cognitive overload. *Cognitive, Affective, & Behavioral Neuroscience*, *7*, 75–89. http://dx.doi.org/10.3758/CABN.7.2.75.

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

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. http://dx.doi.org/10.1016/j.intell.2010.09.001.

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

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology. General*, *133*, 189–217. http://dx.doi.org/10.1037/0096-3445.133.2.189.

Kellogg, C. E., & Morton, N. W. (1999). *BETA III manual*. San Antonio, TX: Harcourt Assessment.

Kiewra, K. A., & Benton, S. L. (1988). The relationship between information processing ability and notetaking. *Contemporary Educational Psychology*, *13*, 33–44. http://dx.doi.org/10.1016/0361-476X(88)90004-5.

Krumm, S., Schmidt-Atzert, L., Buehner, M., Ziegler, M., Michalczyk, K., & Arrow, K. (2009). Storage and non-storage components of working memory predicting reasoning: A simultaneous examination of a wide range of ability factors. *Intelligence*, *37*, 347–364. http://dx.doi.org/10/1016/j.intell2009.02.003.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*, 389–433. http://dx.doi.org/10.1016/S0160-2896(05)80012-1.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, *56*, 1479–1498. http://dx.doi.org/10.2307/1130467.

Martínez, K., Burgaleta, M., Román, F. J., Escorial, S., Shih, P. C., Quiroga, M.Á., et al. (2011). Can fluid intelligence be reduced to 'simple' short-term storage? *Intelligence*, *39*, 473–480. http://dx.doi.org/10.1016/jintell.201.09.001.

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

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive *Gf–Gc* framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 131–150). New York: Guilford.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge, UK: Cambridge University Press.

Moody, D. E. (2009). Can intelligence be increased by training on a task of working memory? *Intelligence*, *37*, 327–328. http://dx.doi.org/10.1016/j.intell.2009.04.005.

Okagaki, L., & Frensch, P. A. (1994). Effects of video game playing on measures of spatial performance: Gender effects in late adolescence. *Journal of Applied Developmental Psychology*, *15*, 33–58. http://dx.doi.org/10.1016/0193-3973(94)90005-1.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology*, *55A*, 1339–1362. http://dx.doi.org/10.1080/02724980244000099.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159–182. http://dx.doi.org/10.1037/0033-295X.111.1.159.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*, 1–48. http://dx.doi.org/10.1111/j.1745-3984.1989.tb00314.x.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Advanced progressive matrices: 1998 edition*. San Antonio, TX: Harcourt Assessment.

Schweizer, K., Goldhammer, F., Rauch, W., & Moosbrugger, H. (2007). On the validity of Raven's matrices test: Does spatial ability contribute to performance? *Personality and Individual Differences*, *43*, 1998–2010. http://dx.doi.org/10.1016/j.paid.2007.06.008.

Shepard, R., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703. http://dx.doi.org/10.1126/science.171.3972.701.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *The American Journal of Psychology*, *15*, 201–292. http://dx.doi.org/10.2307/1412107.

Sternberg, R. J. (2008). Increasing fluid intelligence is possible after all. *Proceedings of the National Academy of Sciences*, *105*, 6791–6792. http://dx.doi.org/10.1073/pnas.0803396105.

van der Ven, A. H. G. S., & Ellis, J. L. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences*, *29*, 45–64. http://dx.doi.org/10.1016/S0191-8869(99)00177-4.

Vandenburg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*(2), 599–601.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250–270. http://dx.doi.org/10.1037/0033-2909.117.2.250.

WASI manual. (1999). San Antonio, TX: Harcourt Assessment.

Wilhoit, B. E., & McCallum, R. S. (2003). Cross-battery assessment of nonverbal cognitive ability. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 63–78). New York: Plenum.

Wright, R., Thompson, W. L., Ganis, G., Newcombe, N. S., & Kosslyn, S. M. (2008). Training generalized spatial skills. *Psychonomic Bulletin & Review*, *15*, 763–771. http://dx.doi.org/10.3758/PBR.15.4.7.