Published in final edited form as:
J Cogn Enhanc. 2017 December ; 1(4): 491–507. doi:10.1007/s41465-017-0047-y.

The Benefits and Challenges of Implementing Motivational Features to Boost Cognitive Training Outcome

Shafee Mohammed[1], **Lauren Flores**[2], **Jenni Deveau**[2], **Russell Cohen Hoffing**[2], **Calvin Phung**[2], **Chelsea M. Parlett**[1], **Ellen Sheehan**[1], **David Lee**[1], **Jacky Au**[1], **Martin Buschkuehl**[3], **Victor Zordan**[4], **Susanne M. Jaeggi**[1], and **Aaron R. Seitz**[2]

[1]School of Education, University of California, Irvine, Irvine, CA 92617, USA

[2]Department of Psychology, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA

[3]MIND Research Institute, 111 Academy Dr., Suite 100, Irvine, CA 92617, USA

[4]School of Computing, Clemson University, 307 McAdams Hall, Clemson, SC 29634, USA

Abstract

In the current literature, there are a number of cognitive training studies that use N-back tasks as their training vehicle; however, the interventions are often bland, and many studies suffer from considerable attrition rates. An increasingly common approach to increase participant engagement has been the implementation of motivational features in training tasks; yet, the effects of such "gamification" on learning have been inconsistent. To shed more light on those issues, here, we report the results of a training study conducted at two Universities in Southern California. A total of 115 participants completed 4 weeks (20 sessions) of N-back training in the laboratory. We varied the amount of "gamification" and the motivational features that might make the training more engaging and, potentially, more effective. Thus, 47 participants trained on a basic color/ identity N-back version with no motivational features, whereas 68 participants trained on a gamified version that translated the basic mechanics of the N-back task into an engaging 3D space-themed "collection" game (Deveau et al. *Frontiers in Systems Neuroscience*, *8*, 243, 2015). Both versions used similar adaptive algorithms to increase the difficulty level as participants became more proficient. Participants' self-reports indicated that the group who trained on the gamified version enjoyed the intervention more than the group who trained on the non-gamified version. Furthermore, the participants who trained on the gamified version exerted more effort and also improved more during training. However, despite the differential training effects, there were no significant group differences in any of the outcome measures at post-test, suggesting that the inclusion of motivational features neither substantially benefited nor hurt broader learning. Overall, our findings provide guidelines for task implementation to optimally target participants' interest and engagement to promote learning, which may lead to broader adoption and adherence of cognitive training.

Correspondence to: Susanne M. Jaeggi; Aaron R. Seitz.

Compliance with Ethical Standards

Conflict of Interest No other authors declare any conflicts of interest.

Keywords

Gamification; Motivation; Brain training; Cognitive training; Task engagement

Working memory (WM) is an underlying mechanism facilitating many daily activities that require storing and manipulating information—examples include mentally adding your monthly bills, determining which route to take to work based on traffic, weather and other factors, combining the ingredients of a dish in the right order (Miyake and Shah 1999), etc. As such, WM is a vital cognitive ability that is highly predictive of how we learn, problem-solve, pay attention, and even adhere to medication regimens (cf. McVay and Kane 2012; Zheng et al. 2011; Gathercole et al. 2003; Higgins et al. 2007; Insel et al. 2006). Due to its limited capacity, WM is one of the bottlenecks for complex thought as well as daily functioning.

Given its relevance in real-world situations, WM interventions have become increasingly popular. Specifically, N-back training has become a promising WM training protocol, and it has shown training benefits that manifest beyond the WM domain (Au et al. 2015, 2016a, b). However, the outcome of individual studies as well as meta-analyses has been mixed (Weicker et al. 2016; Soveri et al. 2017; Melby-Lervåg and Hulme 2013). These differences and inconsistencies can be attributed to multiple factors that contribute to individual and study level factors, such as population characteristics, baseline abilities, personality, training dosage and quality, motivational factors such as remuneration, and gamification of the training protocols—all of which may play a key role in mediating and moderating the training and transfer effects (see Studer-Luethi et al. 2012; Jaeggi et al. 2014; Katz et al. 2016).

In our previous work, we suggested that training paradigms that incorporate attention and reinforcement, multisensory facilitation, multistimulus training, and other game-design elements can maximize training benefits by reinforcing on-task engagement (Deveau et al. 2015). Applying this engaged learning approach to WM training, especially the N-back training paradigm, might potentially foster larger training gains and transfer benefits. For instance, being able to selectively pay *attention* to relevant information in the presence of distractions is suggested to gate learning. Previous studies proposed that WM capacity is correlated with performance on attention tasks and that WM capacity can be improved with training targeted on attention skills (Shiu and Pashler 1992; Leclercq and Seitz 2012). Likewise, rewards that are coincident with times that learning is desirable (such as successes of memory or when reaching more difficult memory challenges) can cause release of neuromodulatory signals that help drive learning (Seitz and Watanabe 2005). Furthermore, training protocols that incorporate *multiple stimuli* lead to greater transfer of training gains (e.g., Dosher and Lu 1998; Xiao et al. 2008). Research also suggests that coordinated *multisensory* processing can increase engagement and lead to extended training benefits provided the stimuli facilitate each other (Kim et al. 2008). Together, these suggest training may be enhanced by incorporating a diversity of stimuli, engaging attention in discerning targets from distractors, providing ample and targeted rewards, and engaging multiple sensory systems.

Well-designed games are known to support learning since the gamification of the learning tasks increases engagement and motivates the participants to persist even if the task becomes difficult (Green and Seitz 2015). Learning via games is known to be effective due to its many features that go beyond simple task engagement—freedom to interpret, experiment, create, fail and recover from failure, and decision making—all of which are critical competencies that empower the learning process (Pellegrini 1995; Rieber 1996). According to Gee (2009), a "good" game that promotes learning typically involves several factors such as rule-based and goal-directed behavior, sense of power and intimacy, providing learning opportunities, creating opportunities to execute actions and supporting the execution of such action (affordances and effectivity), abstract learning experience, and individualized learning paths.

In their review paper, Shute and Ke (2012) identified seven core elements of good games that facilitate learning: (a) interactive problem solving (e.g., solving quests, clearing stages), (b) goal/rule specificity (e.g., advancing to the next level only after completing a certain number of challenges), (c) adaptive challenges (e.g., matching the difficulty of the game to the players abilities), (d) control (e.g., being able to control the game environment, such as change in game speed and taking a different route to complete the stage), (e) ongoing feedback (e.g., on-screen prompts after leveling up), (f) uncertainty (e.g., surprise/bonus stages and multiple endings), and (g) sensory stimuli (e.g., a compelling story telling with a combination of visual graphics and auditory aids).

Still, the implementation of game elements is not always straightforward. In a recent review paper, Green and Seitz (2015) emphasized the importance of proper game (and study) design to achieve the dual purpose of achieving learning goals and engaging participants. They argue that it is not only important to add multiple motivational elements into a game but it is also important to ensure the game is appropriate for the target audience in terms of content (e.g., controlling the levels of violence), features (e.g., providing appropriate stimuli based on the visual capacity), complexity (e.g., contingent with the perceptual abilities such as visual and auditory capacities), and challenge (e.g., contingent with cognitive abilities such as WM capacity). This may help explain why previous attempts to gamify working-memory training have proven to impair task performance, at least in children (Katz et al. 2014) or have shown no benefits on learning in adults, despite participants reporting increased task engagement and enjoyment (Hawkins et al. 2013). Indeed, it has been argued that "motivational" features might distract from the task-relevant features that promote learning, especially in children (Parish-Morris et al. 2013). We suggest, on the contrary, that the problem is not the inclusion of motivation features, but instead that previous attempts to gamify WM training might have been unsuccessful due to a combination of inadequate game design and short-term training. For example, the game design might have included too many features leading to distraction, especially when participants are still learning how to deal with the task requirements as noted by Katz et al. (2014).

In the current work, we seek to shed light on how task-specific game mechanics might benefit participant engagement and ultimately, learning. First, we test whether gamification of an N-back task enhances engagement, on-task learning, and transfer as compared to a non-gamified N-back task. Second, we investigate whether any training-related improvements might generalize to ecologically valid tasks that might have practical benefits in real-life situations.

Participants were randomly assigned to one of two adaptive N-back tasks played on a tablet computer (cf. Fig. 1)—a simple 2D N-back task using a color identification task that lacked any gamification elements ("Tapback") or an engaging 3D space-themed "collection" game that our laboratory has designed and used in the past which is available on the iTunes store ("Recall the Game"; Deveau et al. 2015). Participants had no history of playing either version of the N-back task prior to enrollment in the study. In both training conditions, participants were required to identify whether the current stimulus matched a stimulus presented N-items before by tapping on the tablet screen. Based on previous work (Hawkins et al. 2013), we hypothesized that the participants who received the gamified N-back version would report more engagement and enjoyment as compared to participants in the non-gamified version. In addition, we hypothesized that these game-design elements would lead to equal or greater learning gain both in terms of training benefits and, potentially, on the transfer measures compared to the non-gamified group, given that the game was designed to address the limitations of previous attempts of gamification (Katz et al. 2014; Hawkins et al. 2013; Parish-Morris et al. 2013) where the motivational features may have served to distract players from the memorization task.

Materials and Methods

Participants

Participants consisted of 127 undergraduate students who were recruited from the University of California Riverside and Irvine campuses (average age = 20.02 years, SD = 1.96, range = 17–30 years, 86 women). The study was approved by the review boards from both sites and participants provided written informed consent. Data were collected over seven academic quarters between fall 2014 and fall 2016. Participants volunteered to participate in a study advertised as a "Brain Training study" via flyers and advertisement on social media and received a monetary bonus of $100–150 (depending on the quarter) for their participation. Participants were randomly assigned to train on one of the two variants of the adaptive N-back task and were included in the final sample if they completed at least 14 out of the 20 training sessions, as well as pre-and post-test assessments. Six participants did not show up for the first day of training, four participants left the training after completing less than four training sessions, and one participant each left the training after completing 7 sessions and 12 sessions, respectively (note that there were no systematic differences in attrition rates between the two groups; seven versus five dropouts). Thus, the final analytical sample consisted of 115 participants; 68 participants trained on the gamified N-back (Recall) paradigm (average age = 20.01 years, SD = 2.30, 37 women) and 47 participants played the non-gamified N-back (Tapback) paradigm (average age = 19.93 years, SD = 1.70, 30 women). Notably, the difference sample size between groups is due to the fact that in some quarters, multiple variants of the Recall game were employed (e.g., swipe control versus tilt control, or using different sound variations); however, there were no systematic differences between these different game conditions and so they were combined for the purpose of this report.

Training Tasks

Tapback—Participants were presented with a four-item stimulus set of colored circles (red, blue, green, and yellow) and were required to tap the screen whenever the current stimulus matched the one that was presented N positions back in the sequence (cf. Fig. 1a). Each stimulus was presented for 3 s. Higher levels of N increase WM load and make the task more difficult, and in our study, the “N-Level” reflected the numeric value of N (e.g., Fig. 1a shows an illustration of a two-back task, i.e., an N-level = 2). The N-level progression was adjusted adaptively based on performance, where consistent accuracy above 85% led to an advancement of “N-level” and consistent accuracy below 70% led to a decrement of N-level. Performance feedback was provided in the form of tones indicating correct and incorrect responses. Each 20-min training session consisted of 8–15 blocks with 20–40 trials per round, of which 30% were targets. The exact number of trials per block varied as a function of the adaptive procedure, that is, whenever a new N-level was reached, the first block consisted of 20 trials, and upon successful completion, participants completed 40 trials at that particular N-level. The dependent variable was the maximum N-level achieved per session.

Recall—In the gamified N-back task (“Recall the Game”), participants experienced a reward-based framework designed to reinforce learning outcomes. While the game was not directly inspired by Shute and Ke’s (2012) framework, it did adhere to many of their principles. The problem that players were supposed to solve was how to escape a hostile alien planet through the use of one of the alien’s ships. To accomplish this task, participants needed to navigate the spaceship through wormholes containing stimulus sets based on the N-back mechanic, while also avoiding obstacles (see Fig. 1b and c). The task was to zap (by pressing a button or the ship) target stimuli (colored shapes) that matched the stimuli occurring N items earlier, while also collecting fuel pods (non-targets). To complete a stage, participants need to collect enough fuel pods (~ 50% of the non-targets). By performing well (i.e., performing consistently above 85% accuracy), participants could advance to a new level (a higher N-level). The game is adaptive both in the N-level as well as the difficulty in navigating the ship (that is, both speed and the navigation challenge of the wormholes are made easier or harder based upon navigational success and N-back performance). The game provides users opportunity for control both in their ability to move the ship around the environment and also in that they could adjust the starting speed of the ship between levels. They also experienced pleasant auditory and visual feedback for correct and incorrect responses, and the overall visual and auditory esthetics of the game were appealing and supported the story.

The game also adhered to approaches motivated from the perceptual learning literature (see Deveau et al. 2015), where a multisensory stimulus set of coupled visual and auditory signals, attention grabbing stimuli “popping-up” at onset, and rewarding feedback for accurate task performance were all purposefully engineered to promote learning. Multisensory stimulus sets were employed with consistent relationships between four stimuli sets on the dimensions of color, shape, and sound (where each color was assigned a unique shape and sound). These were broken up into different level-types that focused on different stimulus sets (e.g., color-sound, color-shape, sound-only, and all signals). Task difficulty was varied adaptively and based on level accuracy (hits, misses, and false alarms), ending speed (presentation and response window rate between stimuli at the end of a level), and navigation (level difficulty based on trial length and number of presented obstacles). Similar to the Tapback condition, participants completed 8–15 blocks with 20–40 trials each per block (30% targets), and the dependent variable was the maximum N level achieved per session.

Outcome Measures

We included several outcome measures representing various cognitive domains to test for differential transfer effects as a function of training type. First, we included a measure to assess near transfer, which is typically observed after WM training (e.g., Au et al. 2016a, b; Jaeggi et al. 2010; Soveri et al. 2017). In addition, we focused on two cognitive domains which have shown to be particularly susceptible to the effects of N-back training; specifically, we included measures of inhibitory control, interference resolution (Soveri et al. 2017; Novick et al. 2014), and visuospatial reasoning (Au et al. 2015; Jaeggi et al. 2014). Previous work has demonstrated that those domains share common variance with the N-back task and that they rely on similar neural networks (Hsu et al. 2017; Jaeggi et al. 2010; Szmalec et al. 2011). Furthermore, an exploratory portion of our study investigated the generalization potential of N-back training to more applied measures. For that purpose, we included two measures that were constructed to reflect everyday challenges faced by typical undergraduates (learning from lectures, math), and in addition, we included a measure to assess delay discounting as a proxy for real-world decision making and impulsivity given that those domains are related to WM (Bickel et al. 2011).

Near Transfer Task

Object N-Back Task: This non-trained variant of the N-back task (Jaeggi et al. 2010; Au et al. 2016a, b) was similar to that used in Au et al. (2016a, b) except that the stimuli were pictures of animals (instead of colors) presented in the center of the screen. Participants were asked to respond as quickly as possible indicating whether or not the currently presented stimulus (i.e., the presented animal) was the same as the one presented N positions before. The stimuli were presented for 500 ms, with an inter-stimulus interval of 2500 ms. We used two levels of N-back difficulty, namely, two-back and three-back, with nine blocks at each level. Each block consisted of 20 + N trials containing six targets each. We took two approaches to increase task complexity/difficulty. First, the stimuli consisted of eight different animals that were chosen to reflect similar perceptual and semantic categories (i.e., crab, lobster, penguin, ibis, kitten, lion, stag, and rhino; colors were mainly brown, red, and white). Second, we varied the amounts of lures (i.e., stimuli that appeared one or three trials back in a two-back task; thus, N − 1 or N + 1 lures) such that in each N-back level, there were three blocks with no lures, three blocks with two lures (one N + 1 lure and one N − 1 lure), and three blocks with six lures (three N + 1 lures and three N − 1 lures). The order of the blocks as well as the position of targets and lures was determined randomly. We used the hit rate minus the false alarm rate (pr; Snodgrass and Corwin 1988), as well as reaction times (medians for correct responses, averaged across participants) as dependent variables.

Far Transfer Tasks

Inhibitory Control and Interference Resolution—Given the shared variance of N-back and various measures of inhibitory control and interference resolution (Hsu et al. 2017; Szmalec et al. 2011), we included two measures to capture training-related improvements in those domains and any differential effects as a function of training type.

AX-CPT: In this task (Barch et al. 1997; as used in Au et al. 2016a, b), participants were presented with a stream of letters presented visually in the center of the screen for 250 ms each, followed by a 1000-ms inter-stimulus interval. Participants were required to respond to each stimulus by pressing a pre-specified key for the trials where the letter "X" followed the letter "A" (AX trials—70% of the total trials) and another key for all other trials (response keys were J and F; counterbalanced across participants). Of particular interest were the trials where the letter "A" was followed by a letter other than "X" (AY trials—10% of the trials), thus, reflecting *reactive control* (Braver 2012; Braver and Barch 2002), as well as "BX" trials (10% of the trials), where any non-A letter was followed by the letter "X," reflecting *proactive control* (Braver 2012; Braver and Barch 2002). The rest of the trials consisted of filler trials, that is, "BY" trials (10% of the trials). After 20 practice trials, participants completed six blocks with 70 trials each, with short breaks in between blocks. The dependent variables consisted of the accuracy in percent as well as reaction times (median; correct responses) for the "AY" trials (reactive control) and BX trials (proactive control).

Deese–Roediger–McDermott Paradigm (DRM): We adapted the common DRM paradigm (Stadler et al. 1999) so that it consisted of a wordlist learning task with immediate and delayed free recall, as well as a recognition task. Participants were presented with five lists of 15 words each shown one at a time in the center of the screen (presentation time, 500 ms; interstimulus interval, 2500 ms). Each list consisted of words that were semantically related to one single word that was never presented (critical lure; e.g., for "bread," the related words presented in the list were butter, sandwich, slice, loaf, etc.). The wordlists including the critical lures were all taken from Stadler et al. (1999), and we used parallel-test versions for the pre- and post-test assessments (counterbalanced across participants). After each list, participants were asked to recall and write down as many words as possible from that particular list (in any order). After a delay of about 40 min, the participants were asked to recall as many words as they could from all five lists. Following the delayed recall task, participants completed the recognition task in which they were presented a series of words and asked to indicate for each word whether it was from any of the study lists or whether it was a new word, as well as how confident they were in their decision using a 4-point Likert scale (1 = definitely a new word, 2 = probably a new word, 3 = probably an old word, 4 = definitely an old word). All 75 words from the study lists were presented in random order, and in addition, all critical lures were presented, as well as 25 randomly selected words out of a list of 75 new words that were never shown before. The 75 new words came from five other DRM lists, that is, they were semantically related, but not to the initial study lists. The words remained on the screen until the participant made a response. The dependent variables were the number of correct responses in the immediate and delayed recall phase, as well as the number of (incorrectly) recalled critical lures. For the recognition task, we used the number of correctly remembered items, reaction times (median; all responses), as well as the familiarity rating for the lure items.

Visuospatial Reasoning—We used four visuospatial reasoning tasks that have shown to be susceptible for transfer effects after N-back training (Jaeggi et al. 2014; see also Au et al. 2015 for a meta-analysis). For each of the four tasks, we used parallel-test versions (counterbalanced) for pre- and post-test assessments.

Space Relations Test: In this task (Bennett et al. 1972; as used in Jaeggi et al. 2014), participants were asked to select the appropriate three-dimensional object out of four alternatives that—unfolded—matched the outlines of a two-dimensional pattern. Participants were given two practice items followed by 17 test items. The dependent measure was the number of items solved correctly within 5 min.

Surface Development Test: In this test (Ekstrom et al. 1976; as used in Jaeggi et al. 2014), participants were shown 2D patterns that would form a 3D shape when folded along the lines. Participants were required to match selected sides of the 3D shape with the ones indicated on the 2D pattern. Each of the 3D–2D pairs had five elements to match, requiring a total of 30 responses. The dependent variable was the number of correct responses provided within 6 min.

Form Board Test: The task consists of four target figures that can each be assembled from a combination of five two-dimensional pieces (Ekstrom et al. 1976; as used in Jaeggi et al. 2014). Participants were required to mark the required and unnecessary pieces to correctly make up the target figure. Each participant completed two practice items before completing 24-item sets consisting of five shapes each, yielding 120 total responses. The dependent variable was the number of correct responses provided in 8 min.

Bochumer Matrizen Test (BOMAT): BOMAT (Hossiep et al. 1999) is a reasoning task that consists of multiple 5 × 3 matrices with patterns with one of them missing. Participants were asked to select the appropriate pattern to complete the missing slot from six answer alternatives. After ten practice trials in which participants received feedback, they were given 25 min to solve as many problems as they could (maximum 27). The dependent variable was the number of correctly solved items.

Applied Assessments—In order to capture generalizing effects to real-world tasks, we implemented a range of assessments that are likely ecologically valid and relevant for a student population, namely, learning from lectures and math. We also included a delay discounting task due to the fact that it has been shown to be susceptible to the effects of WM training, especially in substance users (Bickel et al. 2011).

Learning from Lectures: A novel task was created to capture learning from lectures. We used three 3-min videos that were selected from the NSF IGERT video competition, where each video reported a novel research result to a lay audience and thus met our criteria of novelty and understandability. Participants were shown each of the videos, while listening to the audio via headphones, and they were asked to answer 30 questions about the content of the videos following the presentation. The participants were informed beforehand that they were required to memorize the content and the facts presented in each video to be able to successfully answer the questions. Each participant watched three videos at pre-test and three different videos at post-test (counterbalanced across participants). The dependent variable was the total number of correct responses to the questions across the three lecture videos.

Math Task: In this task (adapted from Park and Brannon 2013, 2014, as used in Au et al. under review), participants were asked to perform relatively simple subtraction or addition problems, each consisting of two or three operands ranging from 11 to 244. Correct answers ranged from 11 to 284. Prior to the task, participants completed four practice trials with feedback, and they were not allowed to move on to the actual task until they completed them correctly. They then practiced typing 20 random numbers displayed on the screen in order to prime their fingers to use the number pad and reduce task-irrelevant variability in reaction time measurements. The actual task consisted of 80 trials, with a brief break in the middle. The problems in each trial were unique and randomly generated for each individual, but constrained such that task difficulty was comparable across participants and sessions. Specifically, trials were fully balanced with respect to the number of addition and subtraction trials, carry operations, borrow operations, and number of operands (two or three). Participants were instructed to respond as quickly and accurately as possible. However, unbeknownst to participants, the task timed out after a generous time limit of 25 min in order to prevent excessive fatigue for slow performers. Eighty-nine percent of participants finished the task within this time limit. Percent accuracy as well as the average time (in seconds) to complete all problems served as dependent variables.

Delay Discounting: This task was adapted from the 27-item monetary forced-choice questionnaire classically used to assess delay discounting (Kirby and Maraković 1996), with the exception that all dollar values were inflated up to 2014 standards. Each question consisted of a choice between a smaller immediate reward or a larger delayed reward. For example, “Would you prefer $83 today or $114 in 61 days.” A detailed explanation of the task and the measurement can be found in Odum (2011). The participants were instructed to indicate which of the choices they would make for each scenario as honestly as they could. The dependent variable was the “k” metric as calculated by Odum (2011). Lower k values represent lower discounting of delayed rewards and thus better ability to delay gratification.

Procedure

After providing informed consent, participants underwent two baseline assessment sessions completed on two separate days. The day 1 assessments lasted approximately 90 min and consisted of BOMAT, Lecture videos, DAT Space Relations, and the N-back task, administered in that order. The day 2 assessments lasted approximately 90 min as well, and participants completed DRM immediate recall, ETS Form Board task, Math task, ETS Surface Development task, Delayed Discounting task, DRM Delayed Recall, DRM recognition task, and AX-CPT, in that order.[1] The order of the assessments was kept consistent across participants since our sample size would provide insufficient power to calculate any potential order effects on performance. After pre-test, participants were randomly assigned to one of the two groups. Participants were asked to come to the laboratory to complete their training for 5 days a week over the course of 4 weeks (20 sessions). Participants in both interventions trained in small groups of about four to eight participants at any given time, and they wore headphones in order to minimize distractions. Participants from both groups conducted their training sessions in the same room but were seated at angles where their screens were not easily viewable to each other. After each training session, participants completed a brief survey where they reported their levels of exerted effort and enjoyment. The survey consisted of two questions: (a) How much did you enjoy the game? and (b) I put a lot of effort into this game. Both questions were answered on a 5-point Likert scale with 1 reflecting least enjoyment/effort and 5 reflecting the most enjoyment/effort. After the intervention period, participants completed the post-test assessments following the same protocol as for the pre-test.

Results

Preliminary Analysis

First, to control for outliers, we used three times median absolute deviation (MAD) to winsorize the data, separately for each group and testing session. MAD is a robust measure to identify the spread of the data (Leys et al. 2013). We calculated MAD by first subtracting the median of a dependent variable from each individual value of this variable. The median of these differences represents the MAD. We calculated MAD for every dependent variable and used it to winsorize the data. Any data point which was three times the calculated MAD above or below the median was replaced with the median value plus three times the MAD value for extremely high scores and the median value minus three times the MAD value for extremely low scores.

Overall, there were less than 1% of the datapoints that were affected by this procedure. The next step in our analysis was to investigate whether there were any baseline differences between the groups on any of the measures. There were no statistically significant group differences in any of the pre-test measures (all $p > 0.2$) except for the untrained two-back task (RT) where the Tapback group had faster reaction times than the Recall group [Tapback—$M = 465$, SD = 101; Recall—$M = 554$, SD = 187; $t(113) = 2.97$, $p = 0.004$] and, similarly, in the untrained three-back task [Tapback—$M = 597$, SD = 123; Recall—$M = 685$, SD = 216; $t(113) = 2.51$, $p = 0.013$]. The participants who dropped out of the study did not differ significantly from the participants who completed the study in any of the baseline assessments (all $p > 0.3$). Participants in the non-gamified group completed an average of 18.88 sessions of training (SD = 1.71), and participants in the gamified group completed an average of 18.83 sessions of training (SD = 2.12). There were no statistically significant group differences in the number of training sessions completed [$t(113) = 0.13$; $p = 0.89$].

[1]We administered an additional Face-Name Recall task; however, due to floor performance and technical difficulties, we did not include this task in any of our analyses.

Training Performance

Both groups significantly improved their performance in the training task. The gamified group ("Recall") improved from an average N-back level of 3.28 (SD = 1.23) in the first two sessions to an average N-back level of 4.60 (SD = 1.12) in the last two sessions ($p < 0.01$; d = 1.128). In contrast, the non-gamified group ("Tapback") improved from an average N-back level of 3.38 (SD = 1.22) in the first two sessions to an average N-back level of 4.03 (SD = 1.08) in the last two sessions ($p < 0.01$; $d = 0.67$). A comparison of the normalized training curves illustrates that the Recall group improved more during training than the Tapback group; however, this difference only emerges after session 4 (see Fig. 2a). Overall, the Recall group showed significantly larger training gains (difference between the first two sessions and the last two sessions) than the Tapback group ($p < 0.001$; $d = 0.76$; see Fig. 2b).

The analysis of the self-reported enjoyment questionnaires revealed group differences in the hypothesized direction: participants who trained on the gamified N-back group ("Recall") reported to enjoy the task more than the non-gamified group [$F(1,113) = 5.27$, $p = 0.023$; $\eta_p^2 = 0.50$]. The participants in the Recall group also reported that they exerted more effort than the non-gamified group [$F(1,113) = 3.93$, $p = 0.05$; $\eta_p^2 = 0.22$; see Fig. 3a, b]. Further analyses revealed that self-reported ratings of effort and enjoyment were strongly correlated [$r(113) = 0.69$, $p < 0.001$]. Furthermore, we correlated the average effort and enjoyment with training gains as a function of group (Fig. 4a, b). In the Recall group, neither effort [$r(66) = -0.09$, $p = 0.23$] nor enjoyment [$r(66) = -0.08$, $p = 0.37$] were correlated with the training gains. In contrast, in the Tapback condition, while effort was not correlated with training gain [$r(39) = -0.03$, $p = 0.44$], enjoyment was positively correlated with training gain [$r(39) = 0.26$, $p = 0.04$], indicating that those who showed higher training gain seemed to have enjoyed the training more.

Transfer Performance

The descriptive pre- and post-test data, the test–retest reliabilities (partial correlations accounting for test version), and the effect sizes (accounting for the correlations between the pre-test and post-test measures [$(\mu_2 - \mu_1)/\mathrm{Sqrt}(\sigma_1^2 + \sigma_2^2 - 2r_{12}\sigma_1\sigma_2)$] are provided in Table 1.

In order to capture near transfer effects using the non-trained N-back task, we calculated analyses of covariance (ANCOVAs) comparing the gamified versus the non-gamified group with pre-test performance as covariate and post-test performance as outcome variables for each of the N-back levels separately (pr and RT). We found no significant group effects in any of the N-back measures (all $p > 0.2$, all $\eta_p^2 < 0.2$), except for the two-back RT [$F(2,113) = 3.19$; $p = 0.04$; $\eta_p^2 = 0.18$] (cf. Table 3).

In order to reduce the likelihood of familywise error inflation, we calculated MANOVAs that included a selection of measures based on an exploratory factor analysis on the pre-test scores in order to reduce the number of comparisons. We used the Quartimin oblique rotation technique using the following pre-test measures: BOMAT, DAT Space relationships, ETS Form Board, ETS Surface Development, AXCPT AY trial accuracy, AXCPT BX trial accuracy, DRM free recall falsely remembered items, DRM Recognition familiarity rating (lure items), Math task (accuracy), Lecture videos task (accuracy), and Delay discounting (k). We refrained from including the untrained N-back task since we were interested to understand the impact of training on the far transfer measures. Furthermore, to avoid the issues with dependency and very high correlation between the variables, we focused on accuracy measures, and thus, we did not include the reaction times for math, DRM, and AXCPT. We also excluded the DRM free recall accuracy and DRM recognition accuracy in favor of the variables of most interest, that is, DRM free recall (falsely remembered) and DRM recognition familiarity rating (lure items) given that those two measures require inhibitory control and interference resolution, processes we hypothesized to be related to the trained skills. We identified three factors that together explained 41% of the variance (see Table 2). The first factor included all spatial reasoning tasks (BOMAT, DAT space relations, ETS surface development, and ETS form board), as well as the Lecture video task, accounting for 18% of the total variance, which could reflect broader reasoning abilities. The second factor included AX-CPT AYand BX trials, as well as DRM free recall (falsely remembered items) as well as DRM recognition (familiarity rating; lure items), accounting for 12% of the variance, which could reflect general inhibitory control functions. The math and the delay discounting task (k) loaded on to the third factor, accounting for 10% of the total variance and potentially reflecting general numerical skills (see Table 2). Next, we calculated standardized gain scores for each of the standardized factor variables.

For each of the factors, we calculated a one-way multivariate analysis of variance (MANOVA) using the standardized gains for all the variables within the factor using intervention group as between-subject variable. However, for exploratory purposes, we also calculated individual ANCOVAs for each of the measures, comparing the gamified N-back group and the non-gamified N-back group using post-test performance as dependent variable, and pre-test performance as covariate. Our main hypothesis was that the gamified N-back group would generally outperform the non-gamified N-back group.

Our MANOVAs showed that there were no significant group differences in any of the factor scores (all $p > 0.3$, all $\eta_p^2 < 0.1$). Furthermore, there were no group differences in any of the individual ANCOVAs either (all $p > 0.06$, all $\eta_p^2 < 0.2$). We also analyzed additional variables that were not part of the composite scores, but there were no significant group differences in any of those variables either (all $p > 0.2$, all $\eta_p^2 < 0.1$). The results for all individual ANCOVAs are provided in Table 3.

In order to get a better understanding of the relationship between baseline abilities, training gain, and transfer, we calculated multiple regressions for each of the transfer measures using the post-test score as the dependent variable as well as pre-test performance and training gain as predictors for each of the groups separately. In general, pre-test performance was a significant predictor for post-test performance in most of the variables (see Table 4). However, training gain did not consistently predict transfer. Specifically, in the Tapback group, training gain did not predict training outcome in any of the measures. In the Recall group, training gains predicted performance on the following post-tests: two-back (lure accuracy)—$\beta = 0.28$ (0.02), $p = 0.03$; BOMAT—$\beta = 0.23$ (0.19), $p = 0.05$; lecture videos (accuracy)—$\beta = 0.26$ (0.13), $p = 0.05$; and math (accuracy)—$\beta = 0.24$ (0.12), $p = 0.05$. Finally, we correlated overall self-reported engagement/effort with the gain in each of the outcome measures as a function of group; however, there were no significant relationships (all $r < 0.15$ and $p > 0.2$).

Discussion

This study is one of the few attempting to evaluate the potential benefits of integrating motivational features into an N-back training task. Although both N-back groups showed a significant training benefit, the gamified group showed significantly greater improvement in the trained task, which was especially apparent after the first few training sessions. Thus, our results indicate that adding gaming elements can lead to greater engagement and effort by participants and also enhance the training gains.

Interestingly, there were no differences in terms of training performance between the two groups in the first three sessions, suggesting that for shorter interventions, adding motivational features might not be beneficial, or may even be detrimental, as evidenced by previous work (Katz et al. 2014). Katz and colleagues argued that given that their participants were still in the learning phase during those three sessions, the motivational features might have been distracting and, as such, detrimental for learning. In contrast, our sample consisted of younger adults, which could be the reason why the participants did not feel as distracted by the gamified version. Furthermore, we suggest that the explicit game design in Recall helped ensure that motivational features did not distract from the main task and created an experience of increased engagement without harming the learning experience.

While there was no discernible relationship between enjoyment and training outcome in the Recall group, enjoyment might be associated with training gain in the Tapback group indicating that for some participants who may be driven by intrinsic motivation, gamification might not be a necessary ingredient for enjoyment and, ultimately, learning. However, the observed association between the training gain and the enjoyment in the Tapback condition did not align with our hypothesis. As such, it might be indicative of a spurious effect, which is further corroborated by its marginal statistical significance. Alternatively, this finding might also demonstrate that individual differences substantially affect training outcome and requires further exploration. Specifically, a lack of enjoyment could be detrimental for learning, as we have observed in previous work (Jaeggi et al. 2011, 2014). On the other hand, the absence of a relationship between enjoyment and training outcome in the Recall group suggests that gamification could be a means to address those individual differences and promote learning for all.

Despite the differences in training performance, gamification did not lead to any group differences in the outcome measures, except for two-back RT. However, this group difference is likely driven by a regression to the mean phenomenon (see descriptive measures in Table 1). In comparison with the non-gamified group, the gamified group was subjected to a much broader set of carefully implemented sensory stimuli which were expected to have summative benefits to the learning beyond the training itself. Despite our hypothesis that such complex yet carefully implemented gamification features would boost learning, our data did not provide supporting evidence. Even though our sample size (N =

115) is considerably larger than that of many cognitive intervention studies, any group differences might have been too subtle to be detected. Still, an inspection of the effect sizes of the reasoning tasks and applied measures revealed that they were about twice as large in the *non-gamified group* as compared with the gamified group (reasoning measures, 0.18 vs. 0.09; applied measures, 0.14 vs. 0.07; cf. Table 1), which is the opposite of what we would have expected. Despite the higher effort exerted during game play as reported by the gamified group, it might be possible that the Recall game might have been easier than the Tapback game, leading to more enjoyment and more improvement (Lomas et al. 2013); however, as a result, their domain-general WM and cognitive control skills might have been taxed less during training, which might have led to (numerically) less transfer (Jaeggi et al. 2011).

Another possible explanation for the lack of transfer in the Recall condition is that the gamification added distractions (e.g., noise) that, while sufficient to lead to greater improvement on the N-back than the Tapback, may have still interfered with broader learning. The game Recall involves a visual rich display, multiple sound tracks, as well as navigation challenges that can potentially interfere with improving memory per se. While we have made progress in achieving a game that leads to greater task learning, and equivalent transfer of learning to the non-gamified variant (Katz et al. 2014), further effort is likely required to achieve a game that fully achieves our goal of boosting transfer of learning. It might also be possible that in order for the gamified group to outperform the non-gamified group at post-test, the outcome measures would have to be gamified as well. Specifically, while the post-test was just another session for the non-gamified group, the group that trained on the gamified version might not have exerted their full engagement and effort during post-test given that they might have gotten used to the gamified environment (Murayama et al. 2010). Nonetheless, while training gain did not predict transfer in the Tapback group, there were a few outcome measures in the Recall group that were predicted by training gain, namely, lure accuracy on the two-back task, BOMAT score, accuracies on the math task, and lecture videos. This might be an indication that gamification of the N-back task might have had an impact on the performance of the participants on these tasks, at least to a certain degree.

It is of note that the effect sizes for the various transfer measures are fairly small, which might be related to the fact that participants did not improve as much during training as what we have seen in previous work (e.g., Jaeggi et al. 2008, 2010, 2014). Specifically, in contrast to what we have observed in young adults previously, training performance seemed to have reached ceiling after about six sessions of training.

Overall, despite the promising effects during training, our gamified N-back task is a first prototype and might not yet have the optimal recipe for balancing task enjoyment and on-task learning that is generalizable. Furthermore, participants' self-reports of engagement and enjoyment may not be objective and could be affected by the issues relevant to self-report data (Mitchell 1985). However, given that we had considerable variability in participants' responses, we do not have any reason to believe that there were systematic response biases. Finally, since we did not systematically vary the motivational features in the Recall game in separate conditions, we are unable to indicate which features were more successful than others. Future research needs to address this issue.

Additionally, a key issue is finding an optimal challenge in these games, that is, the adaptive procedures often put participants at levels that are too easy or too difficult, and this may lead to suboptimal learning. This is a particular challenge in a game that is simultaneously adapting on multiple stimulus dimensions. This is further complicated by the fact that individual differences in sensory perception and adaptivity in many dimensions likely render the task easy for certain individuals and challenging for others. Despite the differences between the two training tasks, the training performances presented here (Fig. 2a) are standardized in order to make the comparisons of training performance between the two tasks more meaningful. Seeking to achieve a balance between factors that provide motivation without causing distractions and increase task engagement and enjoyment without having a deleterious effect on the learning process is a sensible direction for future research.

Beyond just searching for the optimal gamification recipe, it is also worthwhile for future research to see if elevated amounts of efforts on the training task are beneficial for long-term transfer despite showing little to no immediate transfer, especially after spaced repetition. After all, most game-related learners benefit due to their long-term commitments to the games, which also entail continuous learning and improvement. One interesting question that we did not further explore in the current work is the role played by gender on the training performance of the participants in the Recall group. It might have been that the cover story and the requirements of the gamified version were not attractive for women, which might have reduced the overall benefits of the training given that the majority of our participants were women. Although our current data do not indicate any significant gender effects on the training performance,[2] further systematic investigations may be needed to understand any potential differences, perhaps by implementing different targeted versions of the cover-story.

While research thus far has shown mixed results as to the benefits of including motivational features in WM training paradigms, our work shows that gamification does increase training performance, suggesting that carefully implemented game design does indeed benefit learning.

Furthermore, gamification does seem to add value for participants in terms of positive training experience and the resulting on-task effort and engagement. The solution to find an optimal set of motivational features that may enhance the WM training benefits to the untrained tasks requires further systematic longitudinal research using even larger samples than the one used here.

[2]Regression analysis were conducted for each group with training gain as the dependent variable and age and gender as the independent variables ($R^2 = 0.003$, $\beta = 0.06$, $p = 0.27$). Furthermore, the correlation between the training gain and enjoyment did not differ by gender in the Recall group ($M = -0.12$ and $F = -0.07$).

Overall Conclusions and Implications

The present study attempted to achieve beneficial use of motivational features incorporated into an N-back training paradigm to understand the training and transfer benefits of such gamification. Given that it was not our aim to test the efficacy of N-back training itself, we did not include an additional control group that did not train on any N-back task here. Prior work has reported inconsistent benefits from gamification of N-back training tasks (Prins et al. 2011; Hawkins et al. 2013; Katz et al. 2014). Results from our study suggest that gamification does indeed lead to benefits in both training engagement and effort but also training performance and, as such, providing evidence for our theoretical account regarding features that make training tasks enjoyable and engaging (Deveau et al. 2014; Green and Seitz 2015). Nonetheless, despite the addition of game-like features to the training task and the resulting differential effects on training outcome, there were very little group differences in the untrained tasks. On the flip side, we did not see any detrimental effects of adding game elements to our training task either.

Overall, it remains difficult to promote participants' best performance via an optimal game design. There are large individual differences between participants that moderate how they interact with the game features that can either add or take away from the overall training efficacy. More research needs to be conducted to overcome these difficulties and to understand how individual task features may mediate training and transfer. We believe that only with an understanding of the factors that moderate and mediate training efficacy can we create an optimal recipe to maximize the outcomes resulting from cognitive training.

Acknowledgments

This work was supported by the National Institute of Health grant no. 1R01MH111742-01 to A.R.S. and S.M.J. M.B. is employed at the MIND Research Institute, whose interest is related to this work, and S.M.J. has an indirect financial interest in the MIND Research Institute.

References

Au J, Sheehan E, Tsai N, Duncan GJ, Buschkuehl M, Jaeggi SM. Improving fluid intelligence with training on working memory: a meta-analysis. Psychonomic Bulletin & Review. 2015; 22(2):366–377. [PubMed: 25102926]

Au J, Buschkuehl M, Duncan GJ, Jaeggi SM. There is no convincing evidence that working memory training is NOT effective: a reply to Melby-Lervåg and Hulme (2015). Psychonomic Bulletin & Review. 2016a; 23(1):331–337. [PubMed: 26518308]

Au J, Katz B, Buschkuehl M, Bunarjo K, Senger T, Zabel C, et al. Enhancing working memory training with transcranial direct current stimulation. Journal of Cognitive Neuroscience. 2016b; 28:1419–1432. [PubMed: 27167403]

Au J, Buschkuehl M, Jaeggi SM. Near and far transfer outcomes of training the approximate number system. (in review).

Barch DM, Braver TS, Nystrom LE, Forman SD, Noll DC, Cohen JD. Dissociating working memory from task difficulty in human prefrontal cortex. Neuropsychologia. 1997; 35(10):1373–1380. [PubMed: 9347483]

Bennett GK, Seashore HG, Wesman AG. Form T, differential aptitude tests, space relations. New York: Psychological Corporation; 1972.

Bickel WK, Yi R, Landes RD, Hill PF, Baxter C. Remember the future: working memory training decreases delay discounting among stimulant addicts. Biological Psychiatry. 2011; 69(3):260–265. [PubMed: 20965498]

Braver TS. The variable nature of cognitive control: a dual mechanisms framework. Trends in Cognitive Sciences. 2012; 16(2):106–113. [PubMed: 22245618]

Braver TS, Barch DM. A theory of cognitive control, aging cognition, and neuromodulation. Neuroscience & Biobehavioral Reviews. 2002; 26(7):809–817. [PubMed: 12470692]

Deveau J, Lovcik G, Seitz AR. Broad-based visual benefits from training with an integrated perceptual-learning video game. Vision Research. 2014; 99:134–140. DOI: 10.1016/j.visres.2013.12.015 [PubMed: 24406157]

Deveau J, Jaeggi SM, Zordan V, Phung C, Seitz AR. How to build better memory training games. Frontiers in Systems Neuroscience. 2015; 8:243. [PubMed: 25620916]

Dosher BA, Lu ZL. Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. Proceedings of the National Academy of Sciences. 1998; 95(23): 13988–13993.

Ekstrom RB, French JW, Harman HH, Dermen D. Manual for kit of factor-referenced cognitive tests. Princeton: Educational Testing Service; 1976.

Gathercole SE, Brown L, Pickering SJ. Working memory assessments at school entry as longitudinal predictors of National Curriculum attainment levels. Educational and Child Psychology. 2003; 20(3):109–122.

Gee JP. Deep learning properties of good digital games: How far can they go? In Serious Games: Mechanisms and Effects. Routledge Taylor & Francis Group; 2009. 67–82.

Green CS, Seitz AR. The impacts of video games on cognition (and how the government can guide the industry). Policy Insights from the Behavioral and Brain Sciences. 2015; 2(1):101–110.

Hawkins GE, Rae B, Nesbitt KV, Brown SD. Gamelike features might not improve data. Behavior Research Methods. 2013; 45(2):301–318. [PubMed: 23055169]

Higgins DM, Peterson JB, Pihl RO, Lee AG. Prefrontal cognitive ability, intelligence, big five personality, and the prediction of advanced academic and workplace performance. Journal of Personality and Social Psychology. 2007; 93(2):298. [PubMed: 17645401]

Hossiep R, Turck D, Hasella M. Bochumer Matrizentest (BOMAT) Advanced. Hogrefe; 1999.

Hsu NS, Jaeggi SM, Novick JM. A common neural hub resolves syntactic and non-syntactic conflict through cooperation with task-specific networks. Brain and Language. 2017; 166:63–77. [PubMed: 28110105]

Insel K, Morrow D, Brewer B, Figueredo A. Executive function, working memory, and medication adherence among older adults. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences. 2006; 61(2):P102–P107.

Jaeggi SM, Studer-Luethi B, Buschkuehl M, Su YF, Jonides J, Perrig WJ. The relationship between N-back performance and matrix reasoning—implications for training and transfer. Intelligence. 2010; 38(6):625–635.

Jaeggi SM, Buschkuehl M, Jonides J, Shah P. Short- and long-term benefits of cognitive training. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(25): 10081–10086. [PubMed: 21670271]

Jaeggi SM, Buschkuehl M, Shah P, Jonides J. The role of individual differences in cognitive training and transfer. Memory & Cognition. 2014; 42(3):464–480. [PubMed: 24081919]

Katz B, Jaeggi S, Buschkuehl M, Stegman A, Shah P. Differential effect of motivational features on training improvements in school-based cognitive training. Frontiers in Human Neuroscience. 2014; 8:242. [PubMed: 24795603]

Katz B, Jones MR, Shah P, Buschkuehl M, Jaeggi SM. Individual differences and motivational effects in cognitive training research. In: Strobach T, Karbach J, editorsCognitive training: an overview of features and applications. Berlin: Springer; 2016. 157–166.

Kim RS, Seitz AR, Shams L. Benefits of stimulus congruency for multisensory facilitation of visual learning. PLoS One. 2008; 3(1):e1532. [PubMed: 18231612]

Kirby KN, Maraković NN. Delay-discounting probabilistic rewards: rates decrease as amounts increase. Psychonomic Bulletin & Review. 1996; 3(1):100–104. [PubMed: 24214810]

Leclercq V, Seitz AR. The impact of orienting attention in fast task-irrelevant perceptual learning. Attention, Perception, & Psychophysics. 2012; 74(4):648–660.

Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology. 2013; 49(4):764–766.

Lomas D, Patel K, Forlizzi JL, Koedinger KR. Optimizing challenge in an educational game using large-scale design experiments. Presented at the CHI; Paris. 2013.

McVay JC, Kane MJ. Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. Journal of Experimental Psychology: General. 2012; 141(2):302. [PubMed: 21875246]

Melby-Lervåg M, Hulme C. Is working memory training effective? A meta-analytic review. Developmental Psychology. 2013; 49:270–291. [PubMed: 22612437]

Mitchell TR. An evaluation of the validity of correlational research conducted in organizations. Academy of Management Review. 1985; 10(2):192–205.

Miyake A, Shah P. Models of working memory: mechanisms of active maintenance and executive control. Cambridge: Cambridge University Press; 1999.

Murayama K, Matsumoto M, Izuma K, Matsumoto K. Neural basis of the undermining effect of monetary reward on intrinsic motivation. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107(49):20911–20916. DOI: 10.1073/pnas.1013305107 [PubMed: 21078974]

Novick JM, Hussey E, Teubner-Rhodes S, Harbison JI, Bunting MF. Clearing the garden-path: Improving sentence processing through cognitive control training. Language, Cognition and Neuroscience. 2014; 29(2):186–217.

Odum AL. Delay discounting: I'm ak, you're ak. Journal of the Experimental Analysis of Behavior. 2011; 96(3):427–439. [PubMed: 22084499]

Parish-Morris J, Mahajan N, Hirsh-Pasek K, Golinkoff RM, Collins MF. Once upon a time: parent–child dialogue and storybook reading in the electronic era. Mind, Brain, and Education. 2013; 7(3): 200–211.

Park J, Brannon EM. Training the approximate number system improves math proficiency. Psychological Science. 2013; 24(10):2013–2019. [PubMed: 23921769]

Park J, Brannon EM. Improving arithmetic performance with number sense training: an investigation of underlying mechanism. Cognition. 2014; 133(1):188–200. [PubMed: 25044247]

Pellegrini AD. The future of play theory: a multidisciplinary inquiry into the contributions of Brian Sutton-Smith. Albany: State University of New York Press; 1995.

Rieber LP. Seriously considering play: designing interactive learning environments based on the blending of microworlds, simulations, and games. Educational Technology Research and Development. 1996; 44(2):43–58.

Seitz A, Watanabe T. A unified model for perceptual learning. Trends in Cognitive Sciences. 2005; 9(7):329–334. [PubMed: 15955722]

Shiu LP, Pashler H. Improvement in line orientation discrimination is retinally local but dependent on cognitive set. Attention, Perception, & Psychophysics. 1992; 52(5):582–588. [PubMed: 1437491]

Shute VJ, Ke F. Games, learning, and assessment. In: Ifenthaler D, Eseryel D, Ge X, editorsAssessment in game-based learning: Foundations, innovations, and perspectives. New York: Springer; 2012. 43–58.

Snodgrass JG, Corwin J. Pragmatics of measuring recognition memory: applications to dementia and amnesia. Journal of Experimental Psychology General. 1988; 117(1):34–50. [PubMed: 2966230]

Soveri A, Antfolk J, Karlsson L, Salo B, Laine M. Working memory training revisited: a multi-level meta-analysis of N-back training studies. Psychonomic Bulletin & Review. 2017; 24(4):1077–1096. [PubMed: 28116702]

Stadler MA, Roediger HL, McDermott KB. Norms for word lists that create false memories. Memory & Cognition. 1999; 27(3):494–500. [PubMed: 10355238]

Studer-Luethi B, Jaeggi SM, Buschkuehl M, Perrig WJ. Influence of neuroticism and conscientiousness on working memory training outcome. Personality and Individual Differences. 2012; 53(1):44–49.

Szmalec A, Verbruggen F, Vandierendonck A, Kemps E. Control of interference during working memory updating. Journal of Experimental Psychology: Human Perception and Performance. 2011; 37(1):137. [PubMed: 20731517]

Weicker J, Villringer A, Thöne-Otto A. Can impaired working memory functioning be improved by training? A meta-analysis with a special focus on brain injured patients. Neuropsychology. 2016; 30(2):190–212. [PubMed: 26237626]

Xiao LQ, Zhang JY, Wang R, Klein SA, Levi DM, Yu C. Complete transfer of perceptual learning across retinal locations enabled by double training. Current Biology. 2008; 18(24):1922–1926. [PubMed: 19062277]

Zheng X, Swanson HL, Marcoulides GA. Working memory components as predictors of children's mathematical word problem solving. Journal of Experimental Child Psychology. 2011; 110(4): 481–498. [PubMed: 21782198]

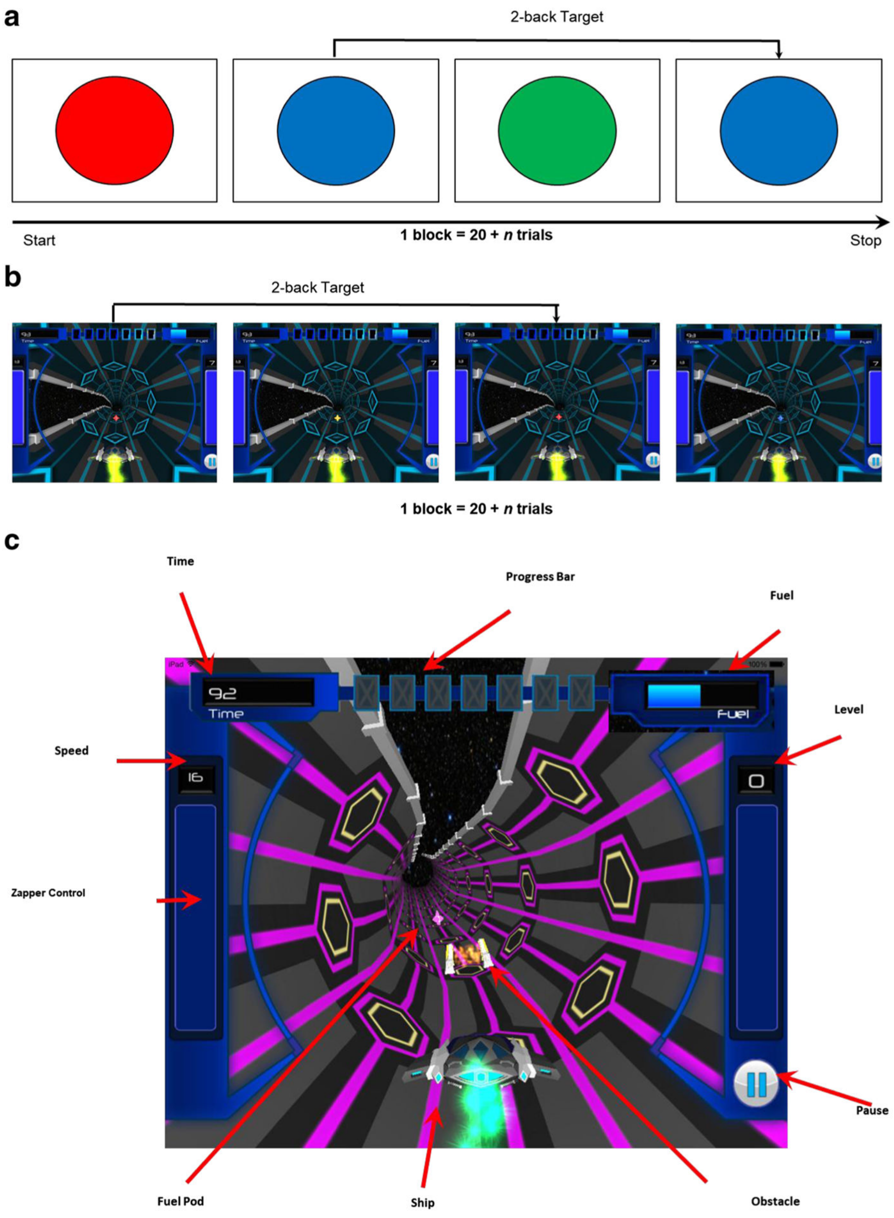

Fig. 1.

a Example for a two-back level in the non-gamified Tapback condition. **b** Example for a two-back level in the gamified Recall condition. **c** Elements of the Recall game that the participant has to monitor. Fuel pod (circled in yellow color) is the target the participant needs to look for

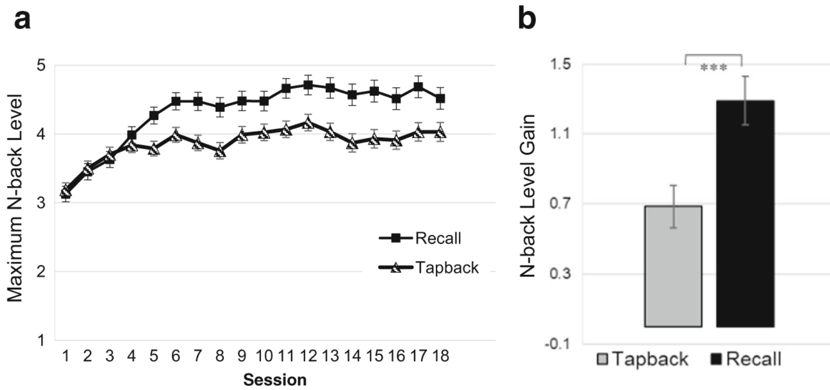

Fig. 2.

a Standardized training performance as a function of session of the participants who completed the training. Error bars represent standard errors. **b** Performance gain in N-back level as a function of group (average performance across the last two training sessions minus average performance across the first two training sessions). Error bars represent the standard errors. ***$p \leq 0.001$

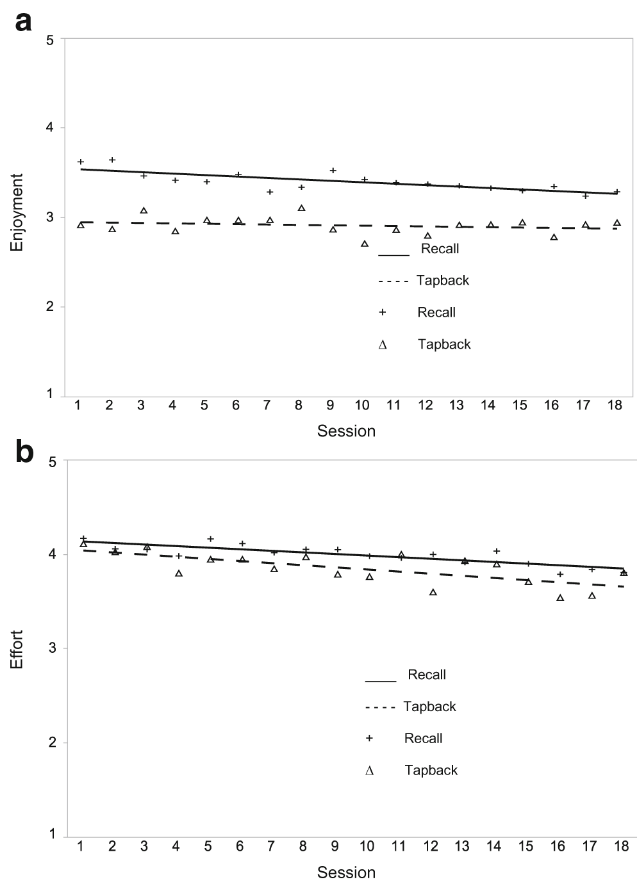

Fig. 3.
a Mean self-reported enjoyment by training group as a function of session. **b** Mean self-reported effort by training group as a function of session

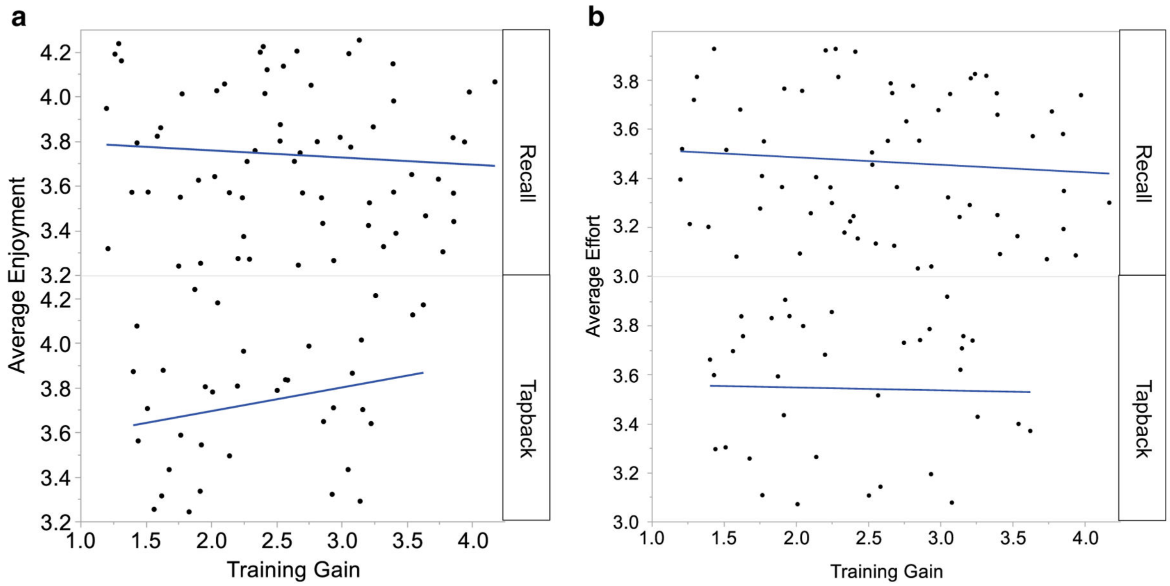

Fig. 4.

a Mean self-reported enjoyment by training gain as a function of group. **b** Mean self-reported effort by training gain as a function of group

Table 1

Descriptive measures and effect sizes for all outcome measures as a function of group

	Pre-test					Post-test								
	N	Mean	SD	Minimum	Maximum	*N*	Mean	SD	Minimum	Maximum	*t*	*p*	*r*	*d*
Tapback (non-gamified N-back)														
Near transfer tasks														
2-Back (pr)	46	0.59	0.22	0.23	0.97	47	0.84	0.13	0.56	1.00	9.02	***	0.62	0.57
3-Back (pr)	46	0.35	0.16	0.13	0.70	47	0.63	0.23	0.15	0.99	10.87	***	0.75	0.58
2-Back (lure accuracy)	46	0.77	0.13	0.50	0.94	47	0.89	0.13	0.50	1.00	5.37	***	0.68	0.42
3-Back (lure accuracy)	46	0.71	0.15	0.33	1.00	47	0.84	0.14	0.50	1.00	4.55	***	0.58	0.41
2-Back (RT, ms)	46	465	101	320	663	47	473	144	182	664	− 0.30		0.73	− 0.03
3-Back (RT, ms)	46	597	123	395	847	47	542	210	264	671	1.53		0.66	0.16
Far transfer tasks														
Inhibitory control and interference resolution														
AXCPT AY trials (accuracy)	47	0.59	0.21	0.21	0.99	47	0.61	0.20	0.19	0.98	0.86		0.82	0.05
AXCPT BX trials (accuracy)	47	0.90	0.07	0.80	1.00	47	0.88	0.08	0.76	1.00	− 1.00		0.67	− 0.13
AXCPT AY trials (RT, ms)	47	407	66	185	532	47	414	101	80	627	− 0.43		0.79	− 0.04
AXCPT BX trials (RT, ms)	47	254	97	135	628	47	243	121	92	669	0.51		0.71	0.05
DRM free recall (accuracy)	47	47.47	7.18	32.18	61.82	47	46.54	9.02	31.00	66.30	− 0.85		0.73	− 0.06
DRM free recall (falsely remembered)	47	3.09	0.90	1.00	4.00	47	1.72	1.38	0.00	3.96	5.82	***	0.61	0.51
DRM recognition (accuracy)	47	31.02	9.03	15.18	44.82	47	30.96	9.66	11.24	51.00	− 0.04		0.72	0.00
DRM recognition familiarity rating (lures)	47	1.59	1.26	0.00	3.96	47	1.78	1.49	0.00	3.96	0.73		0.63	0.07
DRM recognition (RT, ms)	47	1405	333	801	2155	47	1324	283	660	1950	1.27		0.71	0.13
Visuospatial reasoning														
BOMAT	47	13.43	3.00	8.06	19.94	47	14.00	3.07	8.06	19.94	1.28		0.66	0.09
DAT space relationships	47	11.02	3.45	4.94	16.94	47	12.81	3.29	4.10	17.00	4.14	***	0.76	0.26
ETS form board	47	62.96	22.81	20.00	108.00	47	73.60	24.58	25.00	119.00	3.17	***	0.69	0.22
ETS surface development	47	17.40	8.06	7.00	30.00	47	19.79	6.42	6.00	30.00	2.79	**	0.81	0.16
Applied assessments														
Lecture Videos (accuracy)	47	15.70	3.02	9.00	25.00	47	16.28	4.30	9.10	24.00	0.93		0.52	0.08
Math (accuracy)	47	0.85	0.07	0.71	0.97	46	0.86	0.08	0.72	1.00	1.97	*	0.75	0.07

	Pre-test					Post-test								
	N	Mean	SD	Minimum	Maximum	*N*	Mean	SD	Minimum	Maximum	*t*	*p*	*r*	*d*
Math (RT, s)	47	15.55	4.44	10.75	24.59	46	13.24	3.26	9.14	19.94	2.83	**	0.69	0.28
Delay discounting (k)	47	0.02	0.01	0.00	0.04	47	0.04	0.03	0.00	0.09	4.10	***	0.60	− 0.41
Recall (gamified N-back)														
Near transfer tasks														
2-Back (pr)	67	0.64	0.19	0.23	0.93	66	0.82	0.13	0.56	1.00	6.43	***	0.68	0.49
3-Back (pr)	67	0.41	0.18	− 0.01	0.81	66	0.65	0.2	0.3	0.99	7.32	***	0.69	0.53
2-Back (lure accuracy)	67	0.77	0.2	0.11	1.00	66	0.9	0.12	0.46	1.00	4.58	***	0.59	0.37
3-Back (lure accuracy)	67	0.68	0.17	0.22	1.00	66	0.82	0.16	0.22	1.00	4.93	***	0.63	0.39
2-Back (RT, ms)	67	554	187	315	864	66	469	163	189	732	2.81	**	0.73	0.24
3-Back (RT, ms)	67	685	216	455	916	66	623	224	208	761	1.63	0.10	0.69	0.14
Far transfer tasks														
Inhibitory control and interference resolution														
AXCPT AY trials (accuracy)	66	0.65	0.22	0.19	0.98	68	0.66	0.21	0.19	0.98	0.79		0.84	0.02
AXCPT BX trials (accuracy)	66	0.90	0.06	0.8	1.00	68	0.9	0.07	0.76	1.00	− 0.40		0.63	0.00
AXCPT AY Trials (RT, ms)	66	414	60	164	529	68	432	60	237	608	− 1.71	0.07	0.82	0.17
AXCPT BX trials (RT, ms)	66	258	59	141	499	68	238	55	115	398	1.55		0.69	− 0.14
DRM free recall (accuracy)	68	46.66	6.96	32.18	61.82	68	48.48	8.4	33	66.30	1.36		0.64	0.11
DRM free recall (falsely remembered)	68	3.00	0.96	1.00	5.00	68	1.99	1.47	0	3.96	4.72	***	0.69	0.38
DRM recognition (accuracy)	68	29.72	7.76	15.18	44.82	68	32.33	9.54	11.24	52.76	1.74	0.09	0.67	0.15
DRM recognition familiarity rating (lures)	68	1.28	1.13	0.00	3.96	68	1.64	1.33	0.00	3.96	1.69		0.72	0.14
DRM recognition (RT, ms)	68	1497	436	637	2478	68	1348	441	677	2613	1.97	*	0.63	0.17
Visuospatial reasoning														
BOMAT	68	14.09	2.95	8.06	19.94	68	13.97	3.43	8.06	19.94	0.21		0.78	− 0.02
DAT space relationships	68	11.65	3.14	5.06	16.94	68	12.91	3.15	7.00	17.00	2.32	**	0.76	0.20
ETS form board	68	65.38	22.73	27	108	68	72.56	22.07	33	118.00	2.04	*	1.08	0.09
ETS surface development	68	18.56	6.95	5.00	30	68	20.13	7.77	3.00	37.00	1.99	*	0.82	0.09
Applied assessments														
Lecture videos (accuracy)	68	17.54	4.07	10.00	26	68	17.61	4.02	9.10	25.00	0.10		0.69	0.00
Math (accuracy)	68	0.85	0.08	0.71	0.98	68	0.86	0.09	0.72	1.00	0.69		0.63	0.06
Math (RT, s)	68	15.39	4.85	6.63	27.84	68	14.04	4.71	6.73	26.59	1.40		0.69	0.14

	Pre-test					Post-test								
	N	Mean	SD	Minimum	Maximum	*N*	Mean	SD	Minimum	Maximum	*t*	*p*	*r*	*d*
Delay discounting (k)	68	0.02	0.02	0.00	0.04	68	0.03	0.03	0.00	0.09	2.27	**	0.66	− 0.19

All means reflect winsorized values. *r* test–retest reliabilities, *ES* effect size that accounts for correlation between pre-test and post-test measures as in Jaeggi et al. (2014). Numbers of participants varied across testing sessions and variables as a result to loss in data due to technical errors

* $p \leq 0.05$,

** $p \leq 01$,

*** $p \leq 0.001$.

We report any *p* value that is ≤ 0.1

Table 2

Exploratory factor analysis based on the pre-test scores

	Factor		
Pre-test variable	**Reasoning**	**Inhibitory control**	**Numerical skills**
BOMAT	**0.60**	0.06	0.08
DAT space relationships	**0.79**	− 0.05	− 0.02
ETS form board	**0.54**	− 0.04	− 0.15
ETS surface development	**0.64**	0.02	0.18
Lecture videos overall score	**0.37**	0.04	0.09
AXCPT AY trials	0.11	**0.71**	− 0.03
AXCPT BX trials	0.08	**0.57**	0.09
DRM familiarity rating (lures only)	0.06	− **0.17**	− 0.04
DRM falsely remembered	0.11	− **0.47**	0.16
Math	− 0.07	0.01	**0.89**
Delayed discounting (k)	− 0.02	0.00	− **0.13**

N = 115. We used the extraction method of Quartimin oblique rotation technique to derive our factors. Three factors had eigenvalues over 1 and together explained 41% of the variance. Values shown in bold in each column represent the elements that formed a single factor.

Table 3

ANCOVA results for all outcome variables, with post-test as outcome variable, and pre-test as covariate, using group as between-subjects factor

Measure	***F***	***df***	**MSE**	η_p^2	***p* value**
Near transfer tasks					
2-Back (pr)	2.89	2113	0.11	0.19	0.06
3-Back (pr)	0.38	2113	0.18	0.06	0.54
2-Back (lure accuracy)	0.05	2113	0.12	0.02	0.82
3-Back (lure accuracy)	0.75	2113	0.15	0.09	0.39
2-Back (RT)	3.19	2113	188	0.18	0.04
3-Back (RT)	1.97	2113	238	0.14	0.18
Far transfer tasks					
Inhibitory control and interference resolution					
AXCPT AY trials (accuracy)	0.09	2113	0.26	0.07	0.84
AXCPT BX trials (accuracy)	0.70	2113	0.12	0.08	0.40
AXCPT AY trials (RT)	0.51	2113	98	0.07	0.55
AXCPT BX trials (RT)	0.24	2113	111	0.05	0.67
DRM free recall (accuracy)	2.75	2113	7.64	0.17	0.10
DRM free recall (falsely remembered)	0.04	2113	1.12	0.02	0.82
DRM recognition (accuracy)	1.92	2113	8.87	0.14	0.17
DRM recognition familiarity rating (lures)	0.76	2113	1.15	0.09	0.38
DRM recognition (RT)	0.04	2113	406	0.02	0.92
Visuospatial reasoning					
BOMAT	0.51	2113	5.12	0.07	0.47
DAT space relationships	0.32	2113	4.78	0.06	0.57
ETS form board	1.15	2113	35.64	0.11	0.28
ETS surface development	0.57	2113	11.25	0.07	0.45
Applied assessments					
Lecture videos (accuracy)	2.57	2113	4.15	0.16	0.11
Math (accuracy)	0.09	2113	0.15	0.03	0.76
Math (RT)	2.47	2113	3.56	0.16	0.12
Delay discounting (k)	0.35	2113	0.05	0.06	0.56

RT reaction time, *df* degrees of freedom, *MSE* mean squared error

Table 4

Multiple regression analyses for each of the transfer measures, using the post-test score as the dependent variable, as well as pre-test performance and training gain as predictors (analyses are shown separately for each group)

Post-test measure	Beta	SD	*t*	*p* value
Tapback group				
2-Back (pr)				
Pre-test	0.64	0.27	2.37	0.02**
Training gain	0.01	0.03	0.53	0.53
3-Back (pr)				
Pre-test	0.41	0.11	3.63	0.02**
Training gain	0.02	0.02	0.80	0.43
2-Back (lure accuracy)				
Pre-test	0.10	0.17	0.55	0.58
Training gain	0.03	0.02	1.48	0.15
3-Back (lure accuracy)				
Pre-test	− 0.23	0.19	− 1.19	0.24
Training gain	0.03	0.02	1.58	0.12
N-Back (RT)				
Pre-test	0.53	0.12	5.11	0.00***
Training gain	0.00	0.00	0.01	0.99
AXCPT AY trials (accuracy)				
Pre-test	0.64	0.13	5.15	0.00
Training gain	0.00	0.02	− 0.15	− 0.88
AXCPT BX trials (accuracy)				
Pre-test	0.39	0.11	3.33	0.00***
Training gain	0.01	0.01	1.11	0.27
AXCPT AY trials (RT)				
Pre-test	0.42	0.15	4.65	0.00***
Training gain	0.09	0.02	1.00	0.36
AXCPT BX trials (RT)				
Pre-test	0.29	0.17	2.73	0.01**
Training gain	0.06	0.04	1.01	0.32
DRM free recall (accuracy)				
Pre-test	0.47	0.10	4.34	0.00***
Training gain	0.20	0.72	0.28	0.79
DRM free recall (falsely remembered)				
Pre-test	− 0.07	0.07	− 1.07	0.29
Training gain	− 0.01	0.75	− 0.10	0.92
DRM recognition (accuracy)				
Pre-test	0.60	0.14	4.44	0.00***

Post-test measure	**Beta**	**SD**	***t***	***p* value**
Training gain	− 0.17	0.92	− 0.19	0.85
DRM recognition familiarity rating (lures only)				
Pre-test	0.32	0.16	1.94	0.06
Training gain	− 0.08	0.16	− 0.50	0.62
DRM recognition (RT)				
Pre-test	0.34	0.12	14.23	0.00***
Training gain	0.11	0.14	1.27	0.20
BOMAT				
Pre-test	0.48	0.14	3.35	0.00***
Training gain	0.33	0.34	0.98	0.33
DAT space relationships				
Pre-test	0.69	0.11	6.46	0.00***
Training gain	0.09	0.29	0.30	0.77
ETS form board				
Pre-test	0.58	0.13	4.27	0.00***
Training gain	0.09	2.47	0.04	0.97
ETS surface development				
Pre-test	0.85	0.15	5.67	0.00***
Training gain	0.60	0.75	0.80	0.43
Lecture videos (accuracy)				
Pre-test	0.25	0.10	2.62	0.01**
Training gain	0.32	0.32	0.99	0.34
Math (accuracy)				
Pre-test	0.51	0.11	4.48	0.00***
Training gain	0.01	0.01	1.11	0.28
Math (RT)				
Pre-test	0.29	0.09	3.33	0.00***
Training gain	0.01	0.03	1.01	0.33
Delay discounting (k)				
Pre-test	0.27	0.06	4.32	0.00***
Training gain	0.00	0.00	− 0.07	0.94
Recall group				
2-Back (pr)				
Pre-test	0.55	0.19	2.84	0.01**
Training gain	0.02	0.02	0.70	0.49
3-Back (pr)				
Pre-test	0.32	0.13	2.41	0.02**
Training gain	0.01	0.03	0.05	0.64
2-Back (lure accuracy)				
Pre-test	0.36	0.18	2.05	0.05*

Post-test measure	Beta	SD	*t*	*p* value
Training gain	0.05	0.02	2.27	0.03 *
3-Back (lure accuracy)				
Pre-test	0.23	0.18	1.30	0.20
Training gain	0.00	0.03	0.11	0.91
N-Back (RT)				
Pre-test	0.57	0.12	4.04	0.00 ***
Training gain	0.03	0.01	1.01	0.32
AXCPT AY trials (accuracy)				
Pre-test	0.82	0.09	9.31	0.00 ***
Training gain	0.03	0.03	1.48	0.15
AXCPT BX trials (accuracy)				
Pre-test	0.28	0.11	2.49	0.02 **
Training gain	0.00	0.00	− 0.25	0.80
AXCPT AY trials (RT)				
Pre-test	0.66	0.23	4.04	0.00 ***
Training gain	0.07	0.04	0.71	0.48
AXCPT BX trials (RT)				
Pre-test	0.17	0.07	6.54	0.00 ***
Training gain	0.01	0.01	1.01	0.32
DRM free recall (accuracy)				
Pre-test	0.33	0.10	3.21	0.00 ***
Training gain	0.21	0.78	0.26	0.79
DRM free recall (falsely remembered)				
Pre-test	0.12	0.07	1.57	0.12
Training gain	0.05	0.09	0.52	0.61
DRM recognition (accuracy)				
Pre-test	0.39	0.09	4.32	0.00 ***
Training gain	0.61	0.79	0.76	0.45
DRM recognition familiarity rating (lures only)				
Pre-test	− 0.06	0.11	− 0.57	0.57
Training gain	0.06	0.14	0.43	0.67
DRM recognition (RT)				
Pre-test	0.30	0.11	7.03	0.00 ***
Training gain	0.08	0.19	0.43	0.67
BOMAT				
Pre-test	0.38	0.10	3.96	0.00 ***
Training gain	0.59	0.29	2.05	0.05 *
DAT space relationships				
Pre-test	0.62	0.11	5.71	0.00 ***
Training gain	0.27	0.31	0.87	0.39

Post-test measure	**Beta**	**SD**	***t***	***p* value**
ETS form board				
Pre-test	0.54	0.12	4.55	0.00***
Training gain	1.94	2.40	0.81	0.42
ETS surface development				
Pre-test	0.63	0.09	7.33	0.00***
Training gain	0.35	0.60	0.58	0.56
Lecture videos (accuracy)				
Pre-test	0.42	0.12	3.63	0.00***
Training gain	0.88	0.43	2.03	0.05*
Math (accuracy)				
Pre-test	0.33	0.10	3.38	0.00***
Training gain	0.02	0.01	2.01	0.05*
Math (RT)				
Pre-test	0.22	0.11	2.62	0.01**
Training gain	0.03	0.01	0.94	0.36
Delay discounting (k)				
Pre-test	0.29	0.06	4.67	0.00***
Training gain	0.00	0.00	0.17	0.87

All units are standardized

* $p \leq 0.05$,

** $p \leq 0.01$,

*** $p \leq 0.001$