J Cogn Enhanc (2017) 1:539–558
https://doi.org/10.1007/s41465-017-0053-0

Evaluating the Effectiveness of Commercial Brain Game Training with Working-Memory Tasks

Tilo Strobach[1] · Lynn Huestegge[2]

Received: 18 January 2017 / Accepted: 31 October 2017 / Published online: 20 November 2017

Abstract Commercial brain games are home- and computer-based cognitive trainings that are industrially offered and promise to enhance cognitive functioning by repeating cognitive tasks. Despite compelling evidence for the effectiveness of cognitive trainings in various domains and populations, the assumption of brain games' effects on people's minds has been challenged. However, there are only very few attempts to systematically evaluate the effectiveness of such games under ecologically valid training conditions. To approach this gap in the literature, we applied commercially available training tasks assumed to tap into working memory updating and capacity. The effectiveness of this training was measured by utilizing pre- and post-tests in trained tasks (criterion tasks), untrained transfer tasks from the assumed training domains (near-transfer tasks), as well as from the domains processing speed, shifting, inhibition, reasoning, and self-reported cognitive failures (far-transfer tasks). Training as well as pre-post-tests were completely administered home-based. In contrast to an active control group, a training group improved performance in the criterion tasks and near-transfer tasks. Improved performance was also evident in processing speed and shifting tasks (i.e., far-transfer tasks), but these improvements were not as conclusive as those in near-transfer tasks. Further, the number of reported cognitive failures was reduced in the training in contrast to the control group at post-test. Performance improvements were more pronounced for high-performing participants (i.e., magnification effects). In general, this study provides an evaluation of the effectiveness of a particular set of working-memory training tasks in an ecologically valid setting in the context of brain games.

Keywords Commercial brain games · Cognitive training · Transfer, cognitive plasticity · Working memory

✉ Tilo Strobach
tilo.strobach@medicalschool-hamburg.de

[1] MSH Medical School Hamburg, Department of Psychology, Am Kaiserkai 1, 20457 Hamburg, Germany

[2] Department of Psychology, University of Würzburg, Würzburg, Germany

Introduction

"Consumers are told that playing brain games will make them smarter, more alert, and able to learn faster and better. … However, … compelling evidence of general and enduring positive effects on the people's minds … has remained elusive." (A consensus on the brain training industry from the scientific community 2014) (see also Simons et al. 2016).

Brain games, also referred to as brain training, are computer-based cognitive trainings that, broadly defined, aim to enhance a cognitive skill or general cognitive ability by repeating cognitive tasks over a circumscribed timeframe (Rabipour and Raz 2012). Lately, many commercial programs take advantage of this idea over the Internet, offering the comfort and privacy of home-based brain exercise, creating a billion dollar annual industry (Torous et al. 2016). However, there are only very few studies evaluating the effectiveness of commercial home-based cognitive trainings (1) in terms of performance improvements in contrast to a control procedure (see Schmiedek et al. 2010a, for the evaluation of the feasibility of a computer-based cognitive training), (2) under real-life training conditions via the Internet, and (3) in a wider population (for an example of a brain game evaluation exclusively in the elderly, see Nouchi et al. 2012). The present study aims to approach this gap in the literature. An evaluation of training under real-life conditions is required since it allows the evaluation of home-based cognitive trainings with high external validity. That is, real-life training conditions may enable the conclusion that these effects are generalizable to other implementations of computer-based cognitive trainings (Schmiedek 2016).

Previous Studies on Computer-Based Cognitive Trainings

In one of the few existing evaluations, the effectiveness of a computerized, home-based training was investigated using 49 tasks that were presented in game-based formats and were from the various cognitive domains speed of processing, attention, memory, flexibility, and problem solving (Hardy et al. 2015). The study demonstrated that this type of training, when compared with the effects of training of crossword puzzles, led to improvements in core cognitive abilities including speed of processing, working memory, and fluid reasoning. However, the authors of this study had a strong conflict of interest since they were employed by the company that offers the evaluated tasks (see also Nouchi et al. 2012). Further, this study does not allow pinpointing the training domain (or combination of training domains) that is critical for the improvements due to the study's mix of 49 game-based tasks from different domains.

A promising effort in pinpointing the critical training domain was made in a study evaluating the effectiveness of a computerized, home-based training in as many as 11,430 participants (Owen et al. 2010). Recruited among viewers of the BBC science program "Bang goes the theory," participants were assigned to (1) an experimental group with training on the domains reasoning, planning, and problem solving; (2) an experimental group with training on the domains short-term memory, attention, visuospatial processing, and mathematics; as well as (3) a control group with training on answering knowledge questions. Effectiveness of the experimental procedures in contrast to the control procedure was assessed in a pre-post-test design including assessments before the start and after the end of these procedures, respectively. In pre- and post-tests, participants conducted four untrained transfer tasks tapping into (1) reasoning, (2) verbal short-term memory, (3) spatial working memory, and (4) paired-associates learning. In this context, effectiveness is referred to as improved performance in (a set of) untrained transfer tasks to ensure that the improvement is not limited to the trained tasks but generalizes to specific skills and even general cognitive abilities. The results showed that, although improvements were observed in every one of the cognitive tasks that were trained, no evidence was found for transfer effects to untrained tasks, even when trained and untrained tasks were closely related in terms of their underlying cognitive processes. Thus, this evaluation of a computerized, home-based cognitive training (i.e., brain games) showed no effectiveness and thus no positive effects of this type of training.

So, is there no effectiveness of brain games to expect and is the mind immune to changes in cognitive processing with brain game experience? From our perspective, there are at least two lines of reasoning that make these assumptions implausible and call for further investigation of brain games' effectiveness. First, controlled lab-based studies showed that mentally effortful new experiences have the potential to produce changes in those cognitive and neuronal systems that support the acquisition of new skills. For instance, video-game training studies demonstrated positive effects on perception (e.g., Buckley et al. 2010; Green and Bavelier 2007; Li et al. 2009), attention (e.g., Green and Bavelier 2003, 2006a, b; Schubert et al. 2015), processing speed (e.g., Castel et al. 2005; Dye et al. 2009), and executive functioning (e.g., Anguera et al. 2013; Strobach et al. 2012a). Further, there is a large bulk of studies showing such effects as a result of training with working memory tasks (e.g., Au et al. 2015; Klingberg 2010; Koenen et al. 2016; Morrison and Chein 2011; Salminen et al. 2016a, b) (however, for a critical view, see Melby-Lervåg and Hulme 2013; Shipstead et al. 2012) next to demonstrations of positive effects of physical activity, music training, meditation, etc. (for an overview, see Strobach and Karbach 2016).

Second, the study of Owen et al. (2010) shows several limitations in its theoretical framing and methodology; some of them may have made it difficult to draw final conclusions regarding the effectiveness of brain games. For instance, the study provides no theoretical motivation for the selection of training and transfer tasks. That is, while the study lists the sets of these tasks, there is no further elaboration on why these specific sets of training and transfer tasks were selected. This elaboration is essential to allow conclusions about potential cognitive effects and mechanisms of brain games (Green et al. 2014). Another methodological flaw was that the number of training sessions was not controlled in the study of Owen et al., leading to a quite substantial variation between 1 and 188 training sessions across participants. While the relationship between the number of sessions and the changes in the transfer tests was negligible across groups and tests (Owen et al.), there is no report of the relationship between this number and changes in the training tasks. Because of this lacking data report, it is difficult to assess the quality of training effects.

The Present Study

As a consequence of the state-of-the-art literature on brain games, we aimed to test the effectiveness and thus the validity of computer-based cognitive training tasks by targeting a theoretically relevant cognitive structure, namely working memory. Working memory is known as a limited-capacity system that is responsible for the transient storage, processing, and manipulation of information (Baddeley 1986, 2003, 2012; Diamond 2013), and is a system responsible for cognitive processes such as reasoning, decision making, reading comprehension, and regulation of behavior (e.g., Carpenter et al. 1990; Daneman and Carpenter 1980; Engle et al. 1991; Halford et al. 2007). Further, in addition to reports on successful working memory training (e.g., Holmes et al. 2009; Klingberg et al. 2005; Olesen et al. 2004; Westerberg et al. 2007), there is nowadays evidence that working memory training can optimize an individual's performance in a comprehensive range of other cognitive measures, such as cognitive control (e.g., Chein and Morrison 2010; Klingberg et al. 2005), reasoning (e.g., Jaeggi et al. 2008, 2010; Olesen et al. 2004; von Bastian and Oberauer 2013), episodic memory (Dahlin et al. 2008a, b; Richmond et al. 2011; Schmiedek et al. 2010b), and reading comprehension (Chein and Morrison 2010; Karbach et al. 2015). It thus seems plausible to conduct a working memory training to investigate its effectiveness in the context of brain games.

As outlined above, working memory provides both the temporal storage of information as well as the ability to manipulate information (e.g., Baddeley 2012). Thus, generally speaking, working memory is characterized by a structural storage capacity and functional mechanisms that manipulate and update information in this structure (for more specific working memory conceptions, see Oberauer 2009; Wilhelm et al. 2013). From the perspective of this general framework of working memory, our training included tasks assumed to tap into working memory *capacity* and tasks assumed to tap into working memory *updating*. Using both pre- and post-tests, the effects of this training were measured in both trained tasks (criterion tasks) and untrained transfer tasks from the assumed training domains (i.e., capacity and updating in near-transfer tasks). Furthermore, we administered transfer tasks that are structurally dissimilar from the trained tasks (Karbach and Kray 2009; Strobach et al. 2012b). These tasks served to assess far-transfer and thus the selectiveness as well as the effectiveness of the training. More specifically, we considered tasks mainly addressing domains such as processing speed, inhibition, and reasoning. Additionally, we also focused on the assessment of cognitive failures in everyday life. Note that since working memory is a very basic cognitive structure and thus relevant for many cognitive tasks, our categorization of near- vs. far-transfer tasks should not be taken in absolute but relative terms.

From a methodological perspective, we selected and applied training and transfer tasks from a set of tasks that were originally created for a commercially available training program (www.neuronation.com) to maximize external validity (Schmiedek 2016). This selection was exclusively conducted by the authors of the present study and was guided by the distinction into working memory capacity and updating. In detail, we selected tasks that were highly similar to tasks typically utilized in basic cognitive research (in fact, many of the tasks from the commercial program were evidently variants of typical tasks used in basic cognitive research). Also, both training sessions and tasks for the effectiveness (i.e., pre-/post-tests including transfer tasks) were performed home-based (via Internet). This method is equivalent to the study of Owen et al. (2010) and differs from studies combining home-based trainings with lab-based effectiveness tests (e.g., Mahncke et al. 2006; van Muijden et al., 2012), the latter potentially limiting external validity when assessing the effectiveness of an ecologically valid (i.e., real-life) brain game training (Schmiedek 2016; Schmiedek et al. 2010b). Thus, the present study is intended as a controlled large-scale study under real-life training conditions. All participants of the working memory training group received the identical amount of experience with the training tasks, that is, the timing of the tasks was controlled and the training had a fixed extent of 21 sessions.

To control for general effects (i.e., motivational effects, habituation to experimental situation and timed computerized tasks, etc.) as well as test-retest effects (the improvement from pre-test to post-test performance due to the mere repetition of the tasks), we implemented an active control group (note that an exclusive control for test-retest effects would require a passive control group that exclusively perform pre-test and post-test without intermitted training sessions; Green et al. 2014). This active control group's procedure was similar to the procedure in the training group regarding pre-test and post-test sessions, the number of training sessions (thus the amount of training and experience with computerized tasks), as well as the method of home-based training and testing. Further, the active control group performed tasks that were also originally created for a commercially available training program by the company where they signed up for the study, and thus had a very similar layout. Analogous to the training tasks of the training group, the training tasks of the control group included a time-based deadline procedure to keep a constant level of time pressure. In total, the methodological characteristics are comparable between groups, thus also reducing the potential for differences in expectancy effects between groups. The crucial difference to the training group was that the control group was involved in word knowledge tasks tapping into crystallized long-term knowledge, and this group answered knowledge questions about TV news that were shown to the participants. If brain game training has an effect beyond the training tasks, there should be improved performance in near-transfer tasks in the training when compared with the active control group. Further, such improvements in far-transfer tasks would suggest an even broader generalization of brain game training effects, probably due to the essential relevance of working memory for many cognitive requirements.

Methods

Participants

This participants' report has three sections. First, we report characteristics of all participants that initially took part in the study and which were randomly assigned to the two groups (training, control). This set includes those participants that started the study and either dropped out or completed the study within 3 months. Note that participants with a study duration of more than 3 months were excluded from analyses because these participants could not unambiguously be categorized (as participants that dropped out or aimed at completing the program later). Second, we report characteristics of only those participants that completed the study within 3 months. Third, due to still-existing pre-test performance differences between groups (see "Results" section), we excluded further participants and thus report characteristics of those participants that were included in the final matched training and control groups. All participants were paid a 6-month access pass to the brain games offered by the provider NeuroNation after the completion of the study (partial completion of the study was not compensated). Participants were recruited via the website www.neuronation.com and advertisements in various German and Swiss newspapers, internet forums, as well as chats. In these advertisements, potential participants were informed about the general purpose of the training study and the use of the training data for analyses. Participants were randomly and individually assigned to the training and control groups with the requirement to have the same number of participants in each group.

In total, 471 participants started in the study (mean age at start: $M = 41.8$, $SD = 12.4$, range 19–79; 63.7% female) and were randomly categorized as trainees and controls. Then, 86.2% of them completed high school, and 47.6% graduated, respectively. With respect to participants' experience with brain training programs, 11.5% of them had prior experience with brain trainings, while 83.4% were training-naïve (5.1% non-responders). There were 89.0, 7.0, and 4.0% right-handed, left-handed, and ambidextrous participants, respectively, with 95.8% German native speakers; there is no information about the first language(s) of the remaining 4.2%. Table 1 illustrates this information separated for the training and control group.

The group that completed the study included 176 participants with 82 participants in the training group and 94 participants in the control group (mean age at the start of the study: $M = 44.9$, $SD = 11.6$, range 21–78; 64.2% female). From this group, 88.1 and 52.3% completed high school and graduated, respectively. With respect to these participants' experience with brain training programs, 6.3% of them had experience with previous brain trainings, while 83.5% were training-naïve (10.2% non-responders). There were 90.3, 5.1, and 4.5% right-handed, left-handed, and ambidextrous participants, respectively, with 98.3% German native speakers (there is no information about the first language/languages of the remaining 1.7%). Table 1 illustrates this information separated for the training and control group.

The final matched (regarding pre-test performance) sample of training and control groups included 76 participants in each group (mean age in years at the start of the study: $M = 45.1$, $SD = 10.9$, range 21–73; 65.1% females). Further, 86.8 and 52.6% of them completed high school and graduated, respectively. With respect to these participants' experience with brain training programs, 11.2% of them had experience with previous brain trainings, while 82.9% were training-naïve, and 5.9% did not respond. There were 90.8, 4.6, and 4.6% right-handed, left-handed, and ambidextrous participants, respectively, with 98.0% German native speakers (there is no information about the first language(s) of the remaining 2.0%). Table 1 illustrates this information on the matched groups separated for the training and control participants.

General Procedure

The training and control treatment could be conducted on Windows-compatible and Apple-compatible stationary computers and laptops with random-access memory of at least 500 MB. The treatment could be presented on all standard web browsers supporting Java scripts (this excludes Version 9 and lower of the web browser series Internet Explorer as well as all versions of the web browser series Safari); no other technical characteristics required further specification.

In total, the training and control treatment comprised 21 sessions. These sessions were preceded and followed by a pre- and post-test phase, respectively. Sessions during training and pre-/post-test phase lasted approximately 30 and 35 min, respectively. Those training and control participants that completed the study took part in these sessions within 36.5 and 32.5 days, respectively ($t(121.856) = 1.479$, $p > .14$, Levene-test adapted), with all sessions being implemented on separate days.

Pre- and Post-Test Procedure

Both groups performed the identical tasks during their pre- and post-test sessions as listed in Table 2. The order of these tasks was as follows: Missing link, Shuffler, Memory interrupted, Restorer, Turning tops, Turnabout, IQube, d2, Stroop, Digit span, Trail making test (TMT); note that the Cognitive Failures Questionnaire (CFQ; Broadbent et al. 1982) was exclusively conducted during post-test to avoid strong test-retest effects in this test. These sessions included tasks also utilized during training (*criterion tasks*: Shuffler, Memory interrupted), untrained tasks testing training skills (*near-transfer tasks*: Restorer, Turning tops, Turnabout, Digit span), as well as tasks testing untrained skills (*far-transfer tasks*: Stroop, d2, IQube, Missing link, TMT, CFQ). While some of the tasks were taken from the set of commercially available tasks (*www.neuronation.com*), others were specifically designed and included for the purpose of the

Table 1 Overview of the participants information: [1] all participants (that initiated the study), [2] participants that completed the study, [3] final sample of matched participants (in the training and control groups)

	Training group	Control group
[1] All participants		
Female in %	63.6	63.8
Age (Mean/SD/range) in years	42.7/12.5/19–79	40.8/12.2/19–70
Education (high school/university degree) in %	85.2/48.4	87.3/46.6
Experience with braining training (yes/no/no response) in	% 14.4/80.0/5.6	8.1/87.3/4.5
Handedness (right-handed/left-handed/ambidextrous) in %	90.0/6.8/3.2	87.8/7.2/5.0
German native speakers in %	95.6	95.9
[2] Participants that completed the study		
Female in %	65.9	62.8
Age (Mean/SD/range) in years [*]	48.3/11.2/21–78	42.1/11.3/21–71
Education (high school/university)	86.5/59.3	88.5/46.3
Experience with braining training (yes/no/no response) in %	12.3/80.2/7.4	8.4/86.3/5.3
Handedness (right-handed/left-handed/ambidextrous) in %	93.8/2.5/3.7	87.4/7.4/5.3
German native speakers in %	98.8	97.9
[3] Matched participants		
Female in %	64.5	65.8
Age (Mean/SD/range) in years	47.5/10.3/22–73	42.9/11.1/21–71
Education (high school/university)	85.5/61.8	88.2/43.4
Experience with braining training (yes/no/no response) in %	13.2/80.3/6.6	9.2/85.5/5.3
Handedness (right-handed/left-handed/ambidextrous) in %	93.4/2.6/3.9	88.2/66.6/5.3
German native speakers in %	98.7	97.4

present study (implemented with a similar professional design as the other tasks).

Shuffler This task primarily focuses on working memory updating as well as concentration and visuospatial attention. In the beginning of each task trial, participants were presented with, for example, two face-up cards showing objects mixed with three face-down cards (showing a uniform back side) and they were instructed to memorize the object cards. After memorizing, the object cards were turned face-down (showing

Table 2 List of tasks presented during pre-/post-test phase as well as during the treatments of the training and control groups. Tasks during pre-/post-tests are divided into criterion tasks, near-transfer tasks, and far-transfer tasks. Working-memory tasks in the training group are categorized into updating and span tasks, while the control group performed knowledge tasks and answered knowledge questions on TV news (the latter is not listed in the table)

Training tasks in …			
Pre-test	… Training group	… Control group	Post-test
Criterion tasks	Updating tasks	Knowledge tasks	Criterion tasks
Shuffler	Memoflow	Password	Shuffler
Memory interrupted	Parita	Word craft	Memory interrupted
Near-transfer tasks	Memobox	Eloquence	Near-transfer tasks
Restorer	Shuffler		Restorer
Turning tops	Mathrobatics		Turning tops
Turnabout	Span tasks		Turnabout
Digit span	Memory interrupted		Digit span
Far-Transfer tasks	Path Finder		Far-Transfer tasks
Stroop	Reflector		Stroop
d2	Rotator		d2
IQube			IQube
Missing link			Missing link
Trail Making Test 1			Trail Making Test 1
Trail Making Test 2			Trail Making Test 2
			CFQ

their back side) and shuffled with the remaining face-down cards. Finally, a target card from the set of object cards was presented and participants were asked to determine the location of the target card. Task difficulty was adaptive, with adjustments made to the number of face-up cards showing objects as well as the number of face-down cards. After correctly identifying three cards in a row, the difficulty was increased; after two mistakes in a row, difficulty was reduced. Total duration of this task was 2 min.

Memory Interrupted This task is similar to traditional (complex) reading span tasks (Daneman and Carpenter 1980) and taps into working memory span and task shifting. In this task's trials, encoding of memory items (letters or numbers) alternates with processing episodes including the verification of brief mathematical equations such as "2 + 6/2 = 5?". At the end of a trial, participants were instructed to reproduce the list of encoded memory letters and number items. After correctly memorizing all items two trials in a row, difficulty was increased. Difficulty increase means an increase in the processing episode as well as the number of items that participants had to memorize. When participants gave two wrong answers in a row, difficulty was reduced. Total duration of this task was 2 min.

Restorer This task primarily focuses on episodic working memory span. Participants were presented a set of object cards (e.g., illustrations of an apple, ice cream, …) and were instructed to memorize where each object is located in the beginning of each trial. Some of the objects were then hidden and participants should recall the position of each object by clicking on the correct object (choosing from a set of objects at the bottom of the screen). If participants clicked the correct objects two trials in a row, the difficulty was increased and participants had to remember more objects in the following trials. When participants gave two wrong answers in a row, difficulty was reduced. Total duration of this task was 2 min.

Turning Tops This task primarily focuses on working memory updating. At the outset of each trial of this task, participants had to memorize a set of two objects presented on the screen and to click a "Memorized" button. Then, one object disappeared and participants should state whether the remaining object as well as objects of a subsequently presented serial sequence of objects match with the antecessor; thus, this task is similar to a 1-back object task. The position of the object is irrelevant. To respond, participants were instructed to press the left arrow key if the objects match, and the right arrow key if they do not match. Since this exercise is similar to the 1-back paradigm, difficulty was not adaptive. Total duration of this task was 2 min.

Turnabout This task primarily focuses on episodic working memory span. In the beginning of each trial, participants were presented a set of objects in a symmetric grid (e.g., 2 by 2) and were instructed to memorize the objects and their positions. After clicking a "Memorized" button, the objects disappeared and the grid rotated, thus the positions of the objects change. Finally, target objects were shown to the right of the grid and participants had to click the grid's fields where the objects were hidden. If participants clicked the correct objects two trials in a row, the difficulty was increased and participants had to remember more objects in the following trials. When participants gave two wrong answers in a row, difficulty was reduced. Total duration of this task was 2 min.

Digit Span This task primarily focuses on working memory storage capacity and was similar to simple digit-span tasks (Wechsler 2008). In this task's trials, participants were presented with a series of digits (e.g., "8, 2, 4, 3, 7") and asked to immediately reproduce them. If reproduction was successful, they were given a longer list (e.g., "9, 2, 4, 1, 6, 5"). In total, this task contained ten trials with a list length of five in the initial trial and an increase of sequence length every second trial. The relevant dependent variable is the maximal list length successfully reported two times in a row.

Stroop The classic Stroop task (MacLeod 1991; Stroop 1935) taps into the inhibition of prepotent, automatic responses. In this task, colored color words are presented in each trial and participants are instructed to respond via keypress to the word's ink (and ignore the word meaning). In total, a random mix of 30 congruent and 30 incongruent trials were presented with a match or a mismatch between ink and word meaning, respectively. Number of errors and mean RTs in congruent and incongruent trials were recorded, and subtraction (incongruent minus congruent) yielded Stroop effect measures for both RTs and errors.

d2 This neuropsychological test measures selective and sustained attention as well as visual scanning speed (Brickenkamp 1962; Steinborn et al., 2017). In this computerized version of the paper and pencil test, participants are asked to cross out any letter *d* with two marks in total (above and/or below the letter). Surrounding distractors are *p*s or *d*s with 1–4 marks. In sum, there are 14 lines with a total of 47 target and distractor letters each. Time to complete each line is limited to 20 s; the test lasts 4 min and 40 s. This test records the total number of worked-through targets and distractors as well as an error-corrected measure, that is, the number of correctly worked-through items minus the number of incorrectly worked-through items.

IQube This task primarily focuses on logical thinking and reasoning as well as on mental rotation. In each trial, participants were presented with a cube with different, unique shapes at its sides (top of the screen). Participants were instructed to indicate which one of several (rotated) cubes below corresponds to the cube at the top. Response feedback was given afterwards. If participants clicked the correct objects three trials in a row, the difficulty was increased and participants had to choose between a larger number of cubes. When participants gave two wrong answers in a row, difficulty was reduced. Total duration of this task was 2 min.

Missing Link Missing link focuses on logical thinking, reasoning, as well as mental rotation. In each task trial, participants were presented with a sequence of cards with systematically changed shape sets and a final sequence position left empty. At the bottom of the screen, three cards are offered and participants were instructed to indicate the card via the number keypad (i.e., the keys 1 to 3) that logically continues the card sequence. Each trial is finalized with a response feedback. If participants clicked the correct objects seven trials in a row, difficulty was increased and the presented sequence was more complex with respect to the number of stimulus dimensions combined. Total duration of this task was 2 min, thus, the total number of trials could vary across participants.

Trail Making Test The TMT is a neuropsychological test of visual attention and task switching (Arnett and Labovitz 1995). It consists of two parts. The first part requires participants to connect increasing numbers (from 1 to 25) randomly distributed on the screen as quickly as possible while maintaining accuracy (TMT 1). In the second part (TMT 2), participants alternate between connecting numbers and letters (1, A, 2, B, etc.) presented in a similar manner as in the first part. The TMT records the time to complete the first and the second part.

Cognitive Failure Questionnaire The CFQ (Broadbent et al. 1982) assesses the self-report of the frequency of minor everyday lapses (e.g., Do you read something and find you have not been thinking about it and must read it again?). The present version is a computerized German version of the 25-item original test (each item involving a judgment on a scale ranging from 1 to 5, with 5 being equivalent to high cognitive failure proneness), which is also predictive of simple RT-test performance (Steinborn et al. 2016).

Training Procedure

The training group performed tasks that were originally created for commercially available training, provided by www.neuronation.co*m* (Fig. 1a–i). These tasks were working memory updating tasks (i.e., Mixed memories, Parita, Memobox, Shuffler, Mathrobatics) and working memory span tasks (i.e., Memory interrupted, Path finder, Reflector, Rotator), as listed in Tables 2 and 3. The list in Table 3 illustrates which training task was conducted on the first, second, and third set of training sessions (each set of training sessions consisted of seven individual sessions, task order was randomized within each session), and the time period in which each task was trained within the seven-session sets (time in minutes). In total, each task was trained 34 min per participant. As usual in the context of commercially implemented brain trainings, all exercises were adaptive, meaning that the difficulty adjusted to the participants' performance. The operationalization of adaptivity is different for each exercise (see below). The starting difficulty of an exercise on a given training day is always set to the highest difficulty level of all prior training days.

Memoflow This task primarily taps into information processing during working memory updating. In the beginning of each trial, participants were presented with a set of illustrated objects (e.g., an apple) on the left and right of the screen and were instructed to memorize these objects. After clicking a "Memorized" button, the left object disappeared and participants should state whether the remaining right object matches to the left one. If there is a match, participants should click the "same" button; if not, there was no button click. Then, the right object was continuously replaced from trial to trial with the continuous instruction to compare the current with the previous object, similar to an n-back task. If participants responded correctly 19 trials in a row, the difficulty was adapted and the number of objects between the furthermost left and right objects was increased by one. When participants gave two wrong answers in a row, difficulty was reduced. Total duration of this task was 2 min.

Parita The Parita task primarily focusses on working memory updating and multitasking. In the beginning of this task, participants were presented with one visual object and one auditory one-digit number simultaneously. After clicking a "Memorized" button, in the following sequence of trials, participants were instructed to determine whether the auditory number corresponds to the one memorized in the beginning (i.e., similar to a 0-back condition in the context of n-back tasks) and whether the visual object matches the one presented one trial before during the sequence of trials (i.e., similar to a 1-back condition). If either the objects or the numbers are identical, participants should click on a "Sound or image" button on screen; for faster responses, participants were instructed to press the left arrow key. If participants responded correctly 24 trials in a row, difficulty was adapted and the number of objects between the furthermost left and right objects was increased by one. When participants gave two wrong answers in a row, difficulty was reduced. Total duration of this task was 2 min.

Memobox This task taps into working memory processing capacity and visuospatial attention. Trials of this task start with a visual animation of balls entering and leaving boxes.

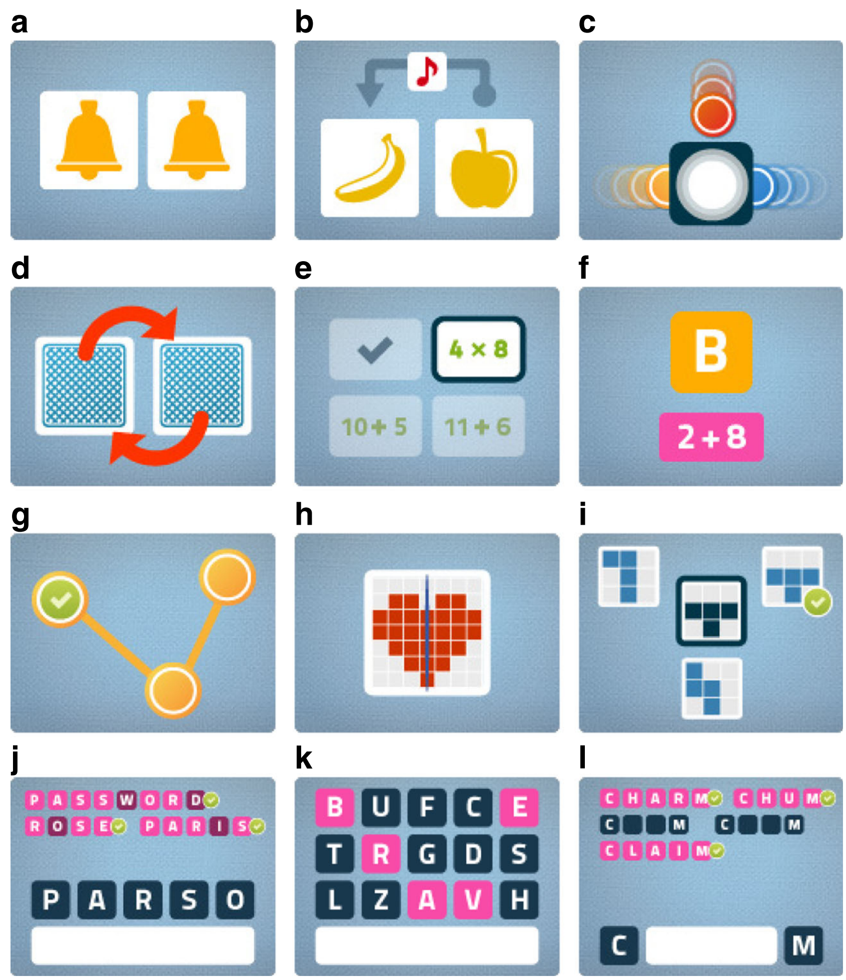


Fig. 1 Illustration of the tasks during the training procedure (**a**–**i**) and the control procedure (**j**–**l**). **a** Memoflow. **b** Parita. **c** Memobox. **d** Shuffler. **e** Mathrobatics. **f** Memory interrupted. **g** Path finder. **h** Reflector. **i** Rotator. **j** Password. **k** Word craft. **l** Eloquence

Participants were instructed to memorize and, subsequentially, to indicate the number and color of balls in each box. If participants responded correctly two trials in a row, difficulty was adapted and the number of balls leaving and entering boxes as well as the number of boxes increased. When participants gave two wrong answers in a row, difficulty was reduced.

Table 3 Overview of training sessions and tasks in the training group. Numbers relate to training time (min)

	Sessions 1–7	Sessions 8–14	Sessions 15–21	Sum
Memoflow		21	13	34
Parita		21	13	34
Memobox		6	28	34
Shuffler		6	28	34
Mathrobatics	20	10	4	34
Memory interrupted	20	9	5	34
Path finder	20	10	4	34
Reflector	20	9	5	34
Rotator	21	10	3	34

Shuffler See pre- and post-test procedure.

Mathrobatics The Mathrobatics task focusses on working memory updating, multitasking, and mental arithmetic. Trials of this task started with the illustration of a mathematical problem and participants were instructed to calculate the result, enter the result using the keyboard or mouse and number pad, as well as to memorize the result. Then participants were instructed to apply displayed numbers for subtraction and addition on the most recently calculated result. The displayed number disappeared and participants had to memorize the result. A second box with a new problem appeared that participants needed to solve, and they had to memorize this result as well. Next, the first box was shown again and they had to continue calculating using the first memorized result. Then the second box appeared again to calculate its mathematical problem using the second memorized result. If participants correctly solved ten problems in a row, the difficulty was adapted and the number of parallel occurring problems as well as the value of the numbers and the difficulty of the operations that had to be computed increased. When participants gave two wrong answers in a row, difficulty was reduced.

Memory Interrupted See pre- and post-test procedure.

Path Finder This task focusses on working memory span and visuospatial attention. At the start of each trial, participants were shown a path of sequentially highlighted dots and participants were instructed to memorize the order in which the dots were connected. After the highlights disappeared, participants were instructed to reproduce the order by clicking the dots in the initially highlighted order (similar to a Corsi block task). If participants responded correctly three trials in a row, the difficulty was adapted and the number of dots increased by one. When participants gave two wrong answers in a row, difficulty was reduced. Total duration of this task was 2 min.

Reflector This task taps into working memory span and concentration. In this task, two parts alternate. In a first part, a grid including a blue field was presented and participants had to memorize the field's position in the grid. After clicking a "Memorized" button, a second part started with participants instructed to decide whether a pattern is vertically symmetrical, horizontally symmetrical, or asymmetrical by clicking onscreen buttons or by pressing corresponding arrow keys. After several alternations of the first and second part, participants were instructed to reproduce the order of the blue fields' grid positions across these alternations. If participants responded correctly two trials in a row, the difficulty was adapted and the number of fields as well as the number of the symmetry-tasks increased. When participants gave two wrong answers in a row, difficulty was reduced. Total duration of this task was 2 min; thus, the total number of trials could vary across participants.

Rotator This mental rotation task taps into working memory span and visuospatial attention. In this task's trials, a set of peripheral patterns is presented around a central target pattern on the screen. Participants are instructed to click the peripheral pattern that is a rotated version of the center pattern. If participants responded correctly four trials in a row, the difficulty was adapted and the number of peripheral patterns as well as their complexity increased. When participants gave two wrong answers in a row, difficulty was reduced. Total duration of this task was 2 min.

Control Procedure

The control group performed three control tasks (Fig. 1j–k) and an additional task involving watching older TV news clips. All tasks were of rather low demand on working memory functions. The tasks the control group received were also selected from the set of tasks provided by www.neuronation.de and thus very similar in overall design, but with a focus on linguistic skills. They were not adaptive, difficulty remained constant irrespective of a participants' performance. Within each of the 20 session during the control procedure, the news issues as well as the other control treatment tasks were ordered in the sequence indicated by the following report.

Watching TV News In each session, participants watched two issues of 20-year-old TV news broadcasts; each issue had a length of 15 min (date of first broadcast is listed per session in Table 6). The broadcasts were originally aired in the time between August and October 1994 on the German broadcasting network ARD. News topics varied from day to day, depending on the newsworthy events. Based on these topics, participants had to give 3–5 multiple-choice responses (four response options each including one correct option). The latter method was implemented to warrant control over participants' attention as well as motivation.

Password Participants were shown strings of letters. They were instructed to find real words containing as many of the letters shown as possible and to enter them via the keyboard. The more letters a word contained, the more points they scored. Total duration of this task was 2 min; thus, the total number of trials could vary across participants. This task primarily focused on verbal fluency.

Word Craft Participants were shown grids of letters. They were instructed to create words out of the given letters (regardless of the letters' location in the grid) as quickly as possible by clicking on these letters. The more letters a word contained, the more points they scored. Total duration of this task was 2 min; thus, the total number of trials could vary across participants. This task mainly focused on visual tracking as well as verbal fluency.

Eloquence Participants were shown the first and the last letter of a word. They were instructed to find the missing letters between the first and last letter of the word. Total duration of this task was 2 min. This task mainly focused on visual tracking as well as verbal fluency.

General Remarks Regarding Tasks in the Pre-/Post-Tests and During Training Note that many of the tasks utilized for the training/control procedures as well as for the pre-/post-tests were adaptive (regarding trial difficulty) and self-paced (regarding trial duration), but characterized by a pre-set time limit (see individual task descriptions), which is typical for commercial brain games. This holds for the tasks Shuffler, Memory interrupted, Restorer, Turning tops, Turnabout, IQube, Missing link, Memoflow, Parita, Memobox, Mathrobatics, Path finder, Reflector, and Rotator. Thus, unlike many lab tasks in basic cognitive research, the number of completed trials within each task and the trial difficulty is not under the control of the experimenter and thus not constant across participants and test time points. Instead, a performance score (arbitrary units) is computed as a dependent variable in many task which reflects both accuracy and speed dimensions of performance. In detail, task difficulty was typically increased either by an increase in processed items (e.g., the number of to-be-memorized items), a shortening of trial duration/increase in frequency of item presentations (the speed dimension), or both. While incorrect responses did not reduce performance scores, correct responses increased these scores and this increase was exponential, that is, the increase was more pronounced at more difficult levels. However, since the algorithms underlying the score computations within each task are the same for all participants and test time points, it is possible to use these scores as performance indices for the sake of comparison across individuals and test time points.

Results

The following computations were executed using the "score" (instead of the alternative "difficulty level") variables for the assessments Shuffler, Memory interrupted, Restorer, Turning tops, and Turnabout; we applied the "score" variable since this variable combines the amount of correct responses, the achieved difficulty level, and elapsed time and thus provides a broader performance measure than simply "difficulty level." For Digit span, we only used the variable indicating the highest number of digits correctly reported two times in a row. For the d2 test, we only used the error-corrected score (note that this measure was very highly correlated with the uncorrected score). The two TMT variables TMT 1 and TMT 2 (reflecting time until completion) were treated separately (i.e., they were not averaged) due to differences in working memory demands between tests (see Methods section). For the Stroop analysis, we computed difference scores (incompatible minus compatible) for both RTs and error rates, resulting in two distinct Stroop effect variables.

Initially, 500 participants were randomly assigned to the 2 groups (training, control). Of these, all 250 participants in the training group but only 221 participants from the control group actually started the pre-test (see "Methods" section). However, only 229/197 participants (training/control) completed all pre-tests (note that participants were not aware of their group assignment at this point). In a first step, we analyzed all available relevant pre-test data using a MANOVA, which included group (training/control) as a factor and 13 dependent variables (Shuffler, Memory interrupted, Restorer, Turning tops, Turnabout, Digit span, Stroop RT effect, Stroop error effect, d2, IQube, Missing link, TMT 1, TMT 2). The MANOVA (Pillai spur used here and throughout the results) revealed a significant group effect, $F(13,412) = 1.744$, $p = .050$, $\eta_p^2 = .052$. Post-hoc ANOVAs revealed significant group differences for the dependent variables memory interrupted ($p < .001$), TMT 2 ($p = .033$), and d2 ($p = .015$), while all other differences were non-significant (ps > .05).

In a second step, we analyzed all participants which contributed complete data sets for pre- and post-tests (82/94 training/control participants). Therefore, we computed a MANOVA with the between-subject factor group and the within-subject factor time (pre-test/post-test) for the 13 dependent variables listed above. This MANOVA resulted in a significant main effect of group, $F(13, 159) = 6.644$, $p < .001$, $\eta_p^2 = .352$, a significant main effect of time, $F(13, 159) = 48.011$, $p < .001$, $\eta_p^2 = .797$, and, importantly, a significant interaction of group and time, $F(13, 159) = 13.954$, $p < .001$, $\eta_p^2 = .533$. Following up on these results, we tested whether the two groups were matched on their pre-test performance regarding the two key variables (i.e., those criterion tasks that were used both during training and for the pre-/post-tests) Shuffler and Memory interrupted. Corresponding independent samples t tests revealed significant pre-test differences for both variables, ts$(174) = 2.60/2.82$, $p = .010/.005$, indicating better overall pre-test performance for the control (vs. training) group and thus unsuccessful group matching.

For a rigorous analysis and interpretation of training effects, it is clearly desirable to have matched pre-test performance between groups. In a third step (and prior to any look at the post-test results), we thus utilized the following matching procedure. As a matching variable, we used the averaged z-standardized pre-test scores for the key variables Shuffler and Memory interrupted (which represented the two targeted training dimensions working memory updating and capacity, see above) and, in a stepwise procedure, deleted the best-performing controls and the worst-performing training participants until the sample consisted of the same number of participants in both groups with equivalent performance on the matching variable. This finally resulted in a sample consisting of 76 controls and 76 training participants with equal pre-test performance. In the matched sample, we did not find a significant difference in the span of training days between the training and the control groups, $t(150) = 1.863$, $p = .065$ (M = 37, SD = 2.5 for training group and M = 32, SD = 1.1 for control group). Thus, we improved control over training in comparison to Owen et al. (2010). During pre-test, we assessed ratings on participants' subjective believe in the effectiveness of brain trainings. A statistical comparison of these quantitative ratings (−1: disbelieve, −0.5: rather disbelieve, 0: undecided, 0.5: rather believe, 1: believe) did not result in a statistically significant group difference, $t < 1$ (overall $M = 0.66$). Concerning the training performance in the tasks that were trained by the control group (i.e., Password, Word craft, Eloquence), we found significantly increased performance from the first training session (after pre-test) to the last training session (before post-test) in these tasks, i.e., Password: $t(75) = 7.71$, $p < .001$, Cohen's $d = 0.88$; Word craft: $t(75) = 3.44$, $p = .001$, Cohen's $d = 0.40$; Eloquence: $t(75) = 8.28$, $p < .001$, Cohen's $d = 0.95$.

Conducting the same two-way MANOVA as referenced above for the final matched sample resulted in similar overall results, including a significant main effect of group, $F(13, 136) = 7.02$, $p < .001$, $\eta_p^2 = .402$, a significant main effect of time, $F(13, 136) = 41.97$, $p < .001$, $\eta_p^2 = .800$, and, importantly, a significant interaction of group and time, $F(13, 136) = 12.38$, $p < .001$, $\eta_p^2 = .542$. Additionally, we aimed to ensure that the significant group*time interaction is not solely driven by the training tasks and thus repeated the same MANOVA without the two training tasks in the statistical model. This analysis yielded the same pattern of results, including the crucial group*time interaction, $F(11, 138) = 4.99$, $p < .001$, $\eta_p^2 = .285$ (note that the same interaction is also significant when analyzing the full, unmatched sample of participants). These results indicate a differential development of performance between groups and thus necessitates further in-depth analyses of the important group*time interaction for the individual dependent variables (univariate tests).[1]

As expected, significant group*time interactions emerged for the two variables that were part of both the pre-/post-tests and the training sessions, Shuffler and Memory interrupted, indicating a training effect, $F(1148) = 22.513$, $p < .001$, $\eta_p^2 = .132$, and $F(1148) = 125.395$, $p < .001$, $\eta_p^2 = .459$, respectively (Fig. 2). Post-hoc contrasts revealed no significant pre-test differences between groups, both Fs < 1, but significant post-test group differences, $F(1, 148) = 17.085$, $p < .001$, $\eta_p^2 = .103$, and $F(1, 148) = 98.139$, $p < .001$, $\eta_p^2 = .399$, respectively.

Furthermore, significant group*time interactions emerged for four transfer variables: Restorer, $F(1148) = 7.818$, $p = .006$, $\eta_p^2 = .050$ (pre-test group difference: $F < 1$, post-test group difference: $F(1148) = 6.624$, $p = .011$, $\eta_p^2 = .043$), Turning tops, $F(1148) = 42.751$, $p < .001$, $\eta_p^2 = .224$, (pre-test group difference: $F < 1$, post-test group difference: $F(1148) = 18.39$, $p < .001$, $\eta_p^2 = .111$), TMT 1, $F(1148) = 5.110$, $p = .025$, $\eta_p^2 = .033$ (pre-post-test differences: $p = .026$ and $p = .347$ for the training and control group, respectively), and TMT 2, $F(1148) = 6.212$, $p = .014$, $\eta_p^2 = .040$ (pre-post-test differences: $p = .319$ and $p = .013$ for the training and control group, respectively). There were no significant interactions regarding the remainder of dependent variables (Turnabout, Digit span, Stroop error effect, Stroop RT effect, d2, IQube, Missing link), all Fs < 1.[2] Figure 2 depicts an overview of the significant findings, while Fig. 4 illustrates the non-significant ones. An independent-samples t test revealed a significant group difference (training: M = 2.03, SD = 0.50, control: M = 2.21, SD = 0.57, $t(140) = 2.034$, $p = .044$) for the CFQ, indicating fewer reported cognitive failure instances for the training group.

Inter-correlations within the training group between all variables (pre- and post-test assessments) that showed significant group*time interactions are depicted in Table 4. None of these variables were significantly correlated with CFQ scores (all ps > .05).

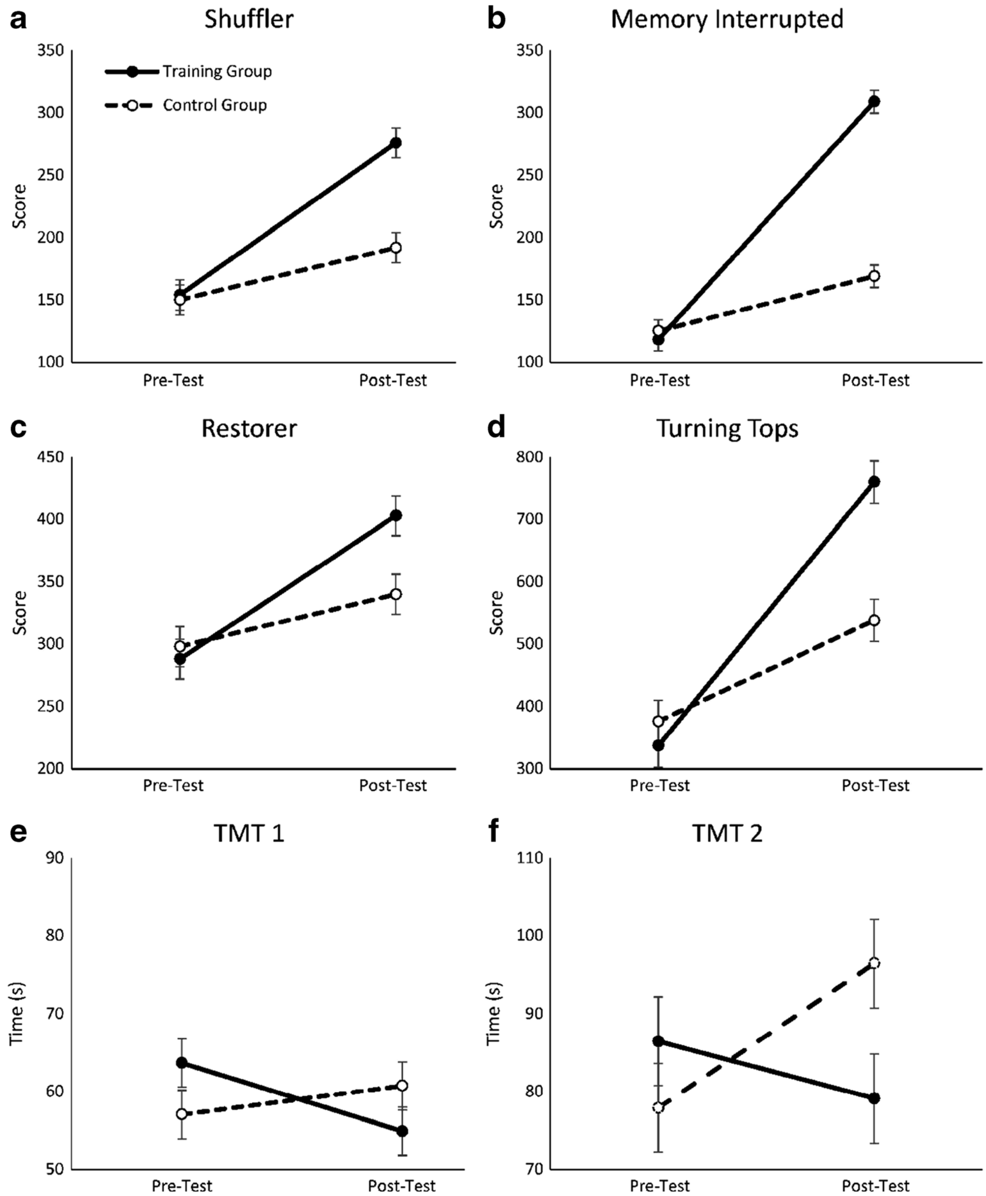
Further, we tested the extent to which the mean individual performance level of participants could predict the training gain (i.e., post-test scores minus pre-test scores) in the training group.

[1] Since the score measures of task performance underlying most of the reports presented here represent rather arbitrary values (reflecting a performance compound related to speed, accuracy, difficulty level, etc.), a probably more psychologically interpretable variable is the difficulty level achieved by participants in each particular task. For example, the difficulty level in working memory tasks is typically defined by the amount of information that needs to be stored and manipulated (objects, numbers etc., see detailed task descriptions in the method section). We therefore additionally ran a MANOVA for the matched sample (excluding the two criterion tasks for a most conservative assessment of transfer effects) in which we used the achieved difficulty level (instead of scores) as a dependent measure for the tasks Restorer, Turning tops, Turnabout, IQube, and Missing Link. The remaining tasks (for which no difficulty level existed as a dependent variable) remained the same as in the previous analysis (Digit span, Stroop error effect, Stroop RT effect, d2, TMT 1, TMT 2). As a result, the MANOVA still revealed the crucial significant group*time interaction, $F(11, 138) = 4.036$, $p < .001$, $\eta_p^2 = 243$. The pattern of significant effects in the subsequent ANOVAs were the same as in the previous analysis: Again, there were significant group*time interactions for Restorer, $F(1148) = 5.612$, $p = .019$, $\eta_p^2 = .037$ (increase in difficulty level from 3.73 to 4.22 (SE = 0.08) in the training group vs. from 3.66 to 3.92 (SE = 0.08) in the control group), Turning tops, $F(1148) = 34.527$, $p < .001$, $\eta_p^2 = .189$ (increase in difficulty level from 1.05 to 2.36 (SE = 0.12) in the training group vs. from 1.20 to 1.67 (SE = 0.12) in the control group), TMT 1, $F(1148) = 5.110$, $p = .025$, $\eta p^2 = .033$, and TMT 2, $F(1148) = 6.212$, $p = .014$, $\eta p^2 = .040$. Thus, the increase in difficulty level from pre- to post-training for the training group was about two times the size of the corresponding increase in the control group for the two variables Restorer (where the difficulty level is directly associated with the number of objects to be remembered) and Turning Tops. There was neither a significant group*time interaction regarding the variable Turnabout, $F(11, 138) = 3.443$, $p = .066$, $\eta_p^2 = .023$, nor regarding the remainder of dependent variables (Digit span, Stroop error effect, Stroop RT effect, d2, IQube, Missing link), all $F < 1$.

[2] To rule out the objection that our final matching procedure affected our results, we additionally followed up on the MANOVA regarding the full sample of participants that contributed pre- and post-test data. These analyses revealed (strikingly similar) post-hoc ANOVA results when compared to the matched sample. Specifically, significant group*time interactions emerged for the two variables that were part of both the pre-/post tests and the training sessions, Shuffler and Memory interrupted, indicating a direct training effect, $F(1171) = 26.554$, $p < .001$, $\eta_p^2 = .134$, and $F(1171) = 141.867$, $p < .001$, $\eta_p^2 = .453$, respectively. Post-hoc contrasts revealed a significant pre-training advantage for the control group ($p = .010$ and $p = .005$), but a significant post-test advantage of the training group ($p = .035$ and $p < .001$, respectively).

Furthermore, significant group*time interactions emerged for four dependent transfer variables: Restorer, $F(1171) = 10.247$, $p = .002$, $\eta_p^2 = .057$ (pre-test advantage of the control group, $p = .009$, but no significant post-test group difference, $p = .359$), Turning tops, $F(1171) = 36.338$, $p < .001$, $\eta_p^2 = .175$ (pre-test advantage of the control group, $p = .022$, post-test advantage of the training group, $p = .014$), TMT 1, $F(1171) = 5.702$, $p = .018$, $\eta_p^2 = .032$ (pre-test advantage of the control group, $p = .008$, no significant post-test group difference, $p = .742$), and TMT 2, $F(1171) = 4.980$, $p = .027$, $\eta_p^2 = .028$ (pre-test advantage of the control group, $p = .003$, no significant post-test group difference, $p = .531$). There were no significant interactions regarding the remainder of dependent variables (Turnabout, Digit span, Stroop error effect, Stroop RT effect, d2, IQube, Missing link), all Fs < 1. Taken together, it is important to note that the final (matched) selection of participants (in order to achieve comparable pre-training performance levels in the criterion tasks between groups) actually worked against our hypothesis of finding significantly greater training effects in the training group. When analyzing the complete set of pre-post data, in many cases a performance disadvantage of the training group in the pre-training assessment was turned into a performance advantage at post-test assessment.

Fig. 2 Overview of training/transfer effects in the matched samples. Performance in the tests Shuffler, Memory interrupted, Restorer, Turning tops, Trail making test (TMT) 1, and TMT 2 is illustrated for pre-test and post-test sessions as well as for the training and the control groups. Note that for TMT 1 and TMT 2, lower durations represent better (faster) performance. Error bars represent SEs



Note that we did not aim to correlate pre-test performance with gain here, since in our data set such a correlation would not yield interpretable results due to statistical artifacts (such as the phenomenon termed "mathematical coupling," see Thorndike 1924; Tu and Gilthorpe 2007), and we also refrained from using Blomqvist's (1977) correction formula for the correlation of pre-test scores and gain, since it requires a reasonable estimate of measurement error based on intra-class correlations for the variable in question (which can only be independently estimated for standardized tests with known re-test reliability). Instead, we correlated the mean of pre- and post-test performance scores (indexing an individual's mean overall ability level in the task) with the training gain (following Oldham 1962). Although there was no significant correlation for the variables Memory interrupted ($r = .150$, $p = .195$) and TMT 2 ($r = .164$, $p = .159$), there were significant correlations for the variables Shuffler ($r = .403$, $p < .001$), Restorer ($r = .392$, $p < .001$), Turning tops ($r = .241$, $p = .036$), and TMT 1 ($r = -.472$, $p < .001$). Thus, most correlations indicate a positive relation between mean overall individual performance levels and the training-related gains, namely that better performance levels are associated with greater training gains (see Fig. 3). There were no significant correlations in the training group between the span of training days and the training gain in any of the variables with a significant group*time interaction, all *p*s > .05.

Finally, we assessed to what extent the *training gain* (in the two training variables Shuffler and Memory interrupted) predicted the *transfer gain* (difference between post- and pre-test scores for all variables displaying significant group*time interactions) in the training group. The results are summarized

Table 4 Inter-correlation matrix of the pre-test (pre) and post-test (post) data of the tests Shuffler, Memory interrupted, Restorer, Turning tops, Trail making test [TMT] 1, and TMT 2 (training group only; $^*p < .05$, $^{**}p < .01$)

	Shuffler (pre)	Memory interrupted (pre)	Restorer (pre)	Turning tops (pre)	TMT 1 (pre)	TMT 2 (pre)	Shuffler (post)	Memory interrupted (post)	Restorer (post)	Turning tops (post)	TMT 1 (post)	TMT 2 (post)
Shuffler (pre)		.176	.187	.383**	−.245*	−.341**	.595**	.252*	.292*	.417**	−.240*	−.332**
Memory interrupted (pre)			.363**	.426**	−.105	−.191	.078	.471**	.274*	.228*	−.051	−.253*
Restorer (pre)				.199	.335**	−.149	.073	.336**	.258*	.212	−.070	−.237*
Turning tops (pre)					−.325**	−.261*	.272*	.260*	.263*	.590**	−.250*	−.183
TMT 1 (pre)						.557**	−.184	−.074	−.006	−.181	.187	.182
TMT 2 (pre)							−.313**	−.155	−.137	−.137	.224	.359**
Shuffler (post)								.398**	.374**	.296**	−.148	−.306**
Memory interrupted (post)									.507**	.366**	−.073	−.362**
Restorer (post)										.243*	.026	−.301**
Turning tops (post)											−.361**	−.261*
TMT 1 (post)												.371**

Note that higher scores are associated with better performance except for the TMT scores (where higher scores indicate lower performance)

in Table 5, and show that the training gain in the Memory interrupted variable significantly predicted the transfer gain for the variables Restorer and Turning tops (r = .250, p = .029 and r = .319, p = .005, respectively), and that the training gain in the Shuffler variable significantly predicted the transfer gain in the Restorer variable (r = .243, p = .035), whereas all remaining associations did not reach statistical significance. In sum, these patterns show a significantly positive relation between training and transfer gains in a way that persons with higher training-related improvement showed higher improvements in the near-transfer tasks.

Discussion

The aim of the present study was to evaluate brain games under training conditions that are as ecologically valid as possible. In detail, we tested the effectiveness of a particular set of computerized home-based training tasks targeting the theoretically relevant cognitive structure working memory and tap into this structure's span component and updating function. The effectiveness of this training was measured by utilizing pre- and post-test assessments in trained tasks (criterion tasks), in untrained tasks from the assumed training domains (near-transfer tasks), as well as in tasks more associated with the domains processing speed, inhibition, task switching, reasoning, and cognitive failures (far-transfer tests).

Discussion of the Present Findings and Relations to Other Training Studies

The comparison of the performance gain between the training and active control groups, which were matched regarding pre-test scores of the two criterion tasks, demonstrated clear training effects. That is, the training group showed a significantly greater performance gain (from pre- to post-test) than the control group. The same result was obtained when all participants that completed the study were analyzed, even though the training group performed worse overall than the control group at pre-test. Importantly, this training effect was consistently demonstrated in both the span and the updating criterion tasks. These training effects in criterion tasks are, for instance, consistent with updating training in other lab-based studies (e.g., Dahlin et al. 2008a, b).

Furthermore, training effects were not limited to the criterion tasks, and the crucial group difference in performance gain was also present when the criterion tasks were not included in the MANOVA. The training group demonstrated advanced performance during the post-test when compared with the control group in the near-transfer tasks Restorer and Turning tops. There is no evidence that this advantage is a result of the pre-test performance, because training and control groups showed similar performance levels at the beginning. These results suggest that the effects of the present home-

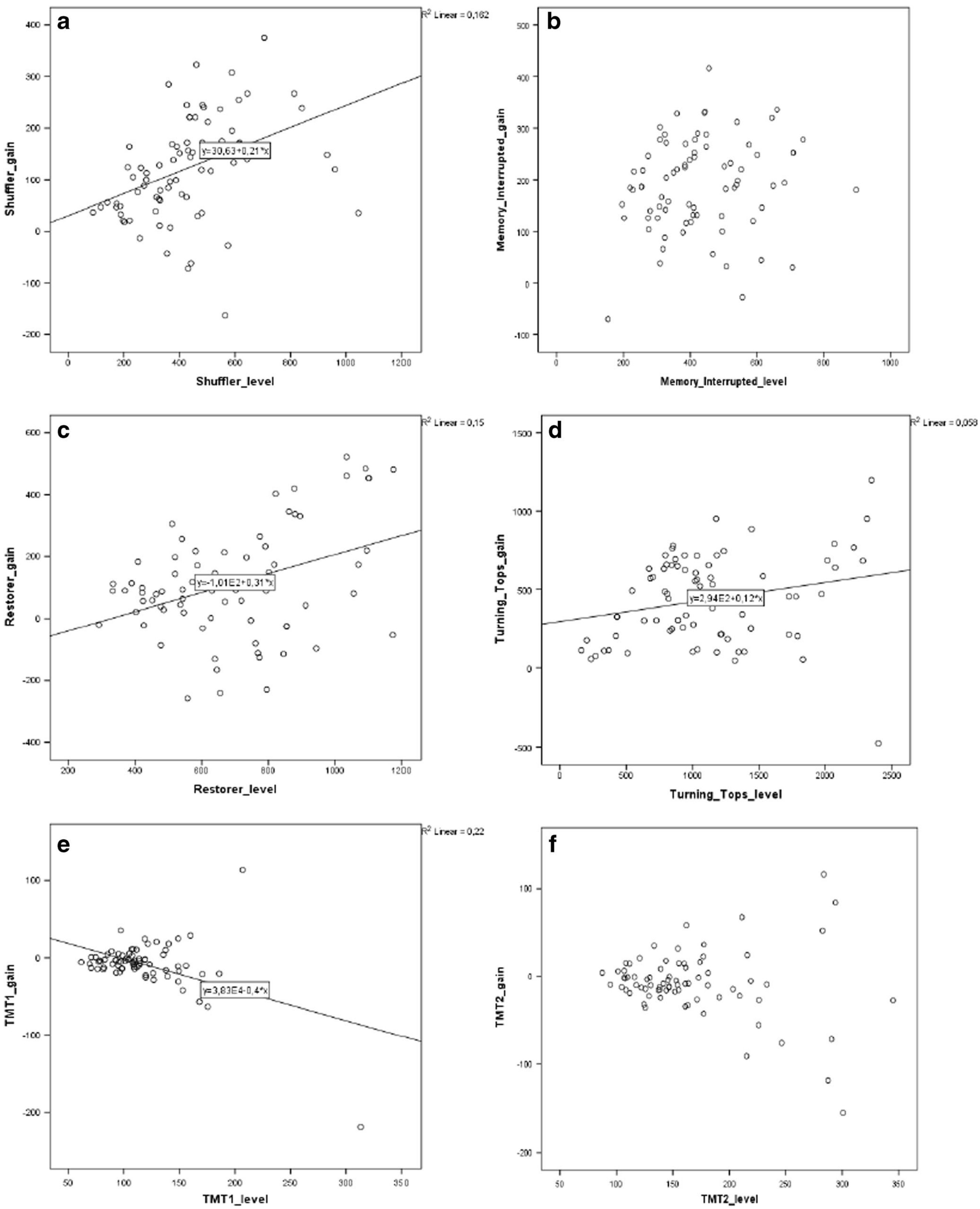


Fig. 3 Scatterplots depicting regressions (training group only) utilizing mean individual performance level (scores averaged across pre-test and post-test) as predictor and individual training/transfer gains as criterion in the tests Shuffler, Memory interrupted, Restorer, Turning tops, Trail making test (TMT) 1, and TMT 2 (Panels A-F, respectively)

Table 5 Results of the regression analyses (multiplicative parameter b, correlation r, and significance index p) with training gain (in the tests Shuffler, Memory Interrupted) as predictor and transfer gain (in the tests Shuffler, Memory interrupted, Restorer, Turning tops, Trail making test [TMT] 1, and TMT 2) as criterion (training group only; * denotes significant correlation, $p < .05$)

Criterion variable (transfer gain)	Predictor variable (training gain)	b	r	p
Restorer	Shuffler	.411	.243*	.035
	Memory interrupted	.482	.250*	.029
Turning tops	Shuffler	.002	.001	.996
	Memory interrupted	.990	.319*	.005
TMT 1	Shuffler	8.81	.027	.815
	memory interrupted	−12.59	.034	.769
TMT 2	shuffler	4.71	.013	.914
	memory interrupted	−53.48	.127	.278

based training are not task-specific effects, but there are effects on non-trained tasks that are assumed to tap into working memory abilities. Such effects of training on transfer tasks are consistent with other lab-based training studies (for an overview see, e.g., Au et al. 2015).

However, based on the present study design, we are not able to disentangle the origin of the effects of training on transfer tasks. It might be that training the updating tasks, the span tasks, or the combination of both task types were responsible for these effects given that all trainees trained both types of tasks. Next, the lacking effects in the near-transfer tasks Turnabout and Digit span are informative because these tasks show that not all types of working memory tasks similarly benefit from the present training. Potentially, Turnabout and Digit span, in comparison to Restorer and Turning tops, share less elements with the training tasks, thus reducing the possibility of transfer of acquired skills (Taatgen 2013) and demonstrating a "task-specificity of effects in near-transfer tasks." More importantly, however, these lacking effects in near-transfer tasks also show that the existence of such effects is no exclusive reflection of improved perceptual processing or motor speed (i.e., trainees simply become faster at pushing buttons). If this was the case, we would expect improved task performance in all tasks from criterion to near and far-transfer tasks, which we did not observe. Thus, the lack of effects in the near-transfer tasks Turnabout and Digit span bears the potential to specify the mechanisms underlying the existing effects (Green et al. 2014).

While there were findings of effects in near-transfer tasks, evidence of far transfer was not as conclusive as effects in near-transfer tasks. That is, we exclusively found evidence for group*session interactions in the tests TMT 1 and TMT 2. In detail, TMT 1 as a standardized test on visual attention showed an improvement in the training group but no such effect in the control group. Thus, it might be that our training improved visual attention. TMT 2 as a task-switching tests provided a similar data pattern: The performance in the training group was improved in contrast to the control group at post-test assessment while there was no such difference at the pre-test assessment. This could be interpreted as evidence that this type of training improved switching abilities and thus cognitive flexibility (Klingberg 2010). This later finding is inconsistent with other studies; these other studies showed no transfer from working memory updating training to the ability to switch between tasks (e.g., Salminen et al. 2016a). However, it should also be noted that the switching demands in the TMT 2 (related to processing alternations between letters and digits) are much less pronounced (and of a somewhat different kind) than in other task switching paradigms (e.g., see Kiesel et al. 2010, for a review), which makes it difficult to directly compare training effects across studies.

However, there are some further issues that need to be discussed in the context of these far-transfer tasks. First, the group*time interaction in TMT 2 might result from performance impairment from pre- to post-test in the control group. Thus, it might be that the control procedure somehow blocked a performance improvement in this particular task (Boot et al. 2011; Kristjánsson 2013) while the training procedure did not. However, if the far-transfer effect in this task is rooted in a reduced blocking effect in the training group, this might be interpreted at least as an indirect proximal frame of evaluation (Schmiedek et al. 2010a), that is, an indirect benefit of the computerized, home-based training. Second, the fact that significant group*time interactions emerged in both TMT versions might indicate that the relative benefit of the training group in TMT 2 is (at least partly) based on skills also relevant for TMT 1. While an analysis using different scores between both TMT measures could at first sight help to isolate the unique switching component in TMT 2, there are also well-known difficulties associated with the correlation of different scores, which are notoriously unreliable (Caruso 2004; Cronbach and Furby 1970; Hedge et al., 2017; Rogosa 1995; Willett 1988). Thus, we consider it premature to draw any final conclusions regarding beneficial training effects on switching skills based on the TMT 2 findings.

At first sight, one might further argue that both TMT versions (i.e., TMT 1 and TMT 2) are somewhat similar to the training task Path finder and might thus not represent far-transfer tasks in the first place. In detail, both tasks are related with processing of subsequently presented points that will be connected. However, despite this superficial similarity, we rather assume that quite different underlying processes are at play. That is, in Path finder, participants were instructed to memorize the order of highlighted dots and were instructed to reproduce the order. In contrast to this task that mainly requires reproduction processes, all points are permanently on screen in the TMT task. Thus, the assumed working memory load should be rather low for TMT which rather focusses on visual attention, while Path finder has a strong working memory component. Therefore, we assume that in terms of underlying processes, the TMT tasks do not resemble the Path finder task, and thus can still be considered tasks testing for far transfer. Nevertheless, we consider it overall premature to draw any final conclusions regarding far-transfer effects since evidence for such effects is not as conclusive as for near-transfer effects, but is rather modest.

Furthermore, we showed evidence for differences between the training and control group in the frequency of self-reported everyday lapses (i.e., CFQ); thus, the training might also affect situations beyond computer-based tests in everyday life. However, one has to consider that, while there is a reduced number of self-reported everyday lapses after training in contrast to the control procedure, there are no corresponding pre-test data. Thus, we cannot directly exclude the possibility that group differences during post-test assessment result from pre-test differences. However, it is important to note that no group comparison at pre-test for other variables provided evidence for beneficial performance in the training (vs. control) group. Therefore, we see no evidence that the CFQ's post-test differences result from strong differences already present prior to training.

On an individual level, our data largely support the assumption of a magnification effect (also called Mathew effect) because the training resulted in a magnification of general individual differences. That is, largest performance improvements in criterion and transfer tasks occurred in trainees with highest average performance. In other words, trainees that needed training the least appeared to have benefitted the most, mirroring a pattern observed in children's reading development (e.g., Stanovich 1986; Walberg and Tsai 1983) and memory-strategy training (Rebok et al. 2007). The magnification effect might be explained by the fact that high-performance trainees build on earlier success, acquired skills, and/or more efficient cognitive resources to become even better performers with the present intervention. This better performance might result from the acquisition and implementation of new strategies and abilities that could not be acquired by persons with low training performance. These observations of a magnification effect clearly speak against the presence of a compensation effect (i.e., that low performing individuals exhibit the greatest training benefits), and are rather inconsistent with similar observations from working-memory training studies and their relevant transfer effects (e.g., Dahlin 2013; Karbach et al. 2015).

Note that our main analysis is based on matched training and control groups. In detail, we excluded those participants from the control group that showed high performance values in the criterion tasks and those participants from the training group with low values. Could this matching procedure artificially produce training and transfer effects in the training in contrast to the control group? We are confident that this is not the case, since the overall pattern of effects in training/transfer tasks remained unchanged in a supplementary analysis involving all available (i.e., non-matched) participants that contributed complete pre- and post-test data (see Footnote 1).

Furthermore, it is essential to ask how participants who completed the study (i.e., pre-test—training/control treatment—post-test) differed from those who did not; this assessment might contribute to our knowledge on drop-outs from online, home-based studies. When analyzing these two groups of participants, we found that they did not differ significantly in the two criterion tasks at pre-test (*t*s < 1). However, they differed significantly regarding Stroop performance ($p < .05$), and on a close to significant level ($p = .065$) regarding digit span performance. Does the CFQ questionnaire as a self-report measure on everyday life inform about potential drop-outs? Unfortunately, we cannot assess the relation between drop-out and CFQ since CFQ was exclusively performed during the post-test assessment. Thus, drop-outs did not perform this questionnaire and a comparison with participants who completed the study was not conceivable. Further, we cannot assess the relation between effects in criterion and transfer tasks between drop-out participants and those that also completed the post-test assessment since the former group did not perform the post-test which is necessary to assess these effects.

Potential Limitations and Open Issues

Concerning potential limitations of our study, it is essential to point out that we are lacking control over the participant's computer, that is, timing, screen, input devices, output devices, the computer's processing characteristics, etc. However, since we run the present evaluation on a variety of computer environments, we assume to have increased the external validity of the present evaluation (Schmiedek 2016). Further, the present study lacks control regarding concurrent actions and tasks of the participants on the computer and beyond. Actions and tasks on the computer could be controlled by installing a program on the local computer (Schmiedek et al. 2010a). However, ethical and privacy issues speak against such a procedure. The control of actions and tasks beyond the computer would require a lab-based evaluation of the present training program, and it is clearly desirable to evaluate the present training program in a controlled lab situation in the future to come up with converging evidence regarding its effectiveness. Such lab-based evaluations could also include more standardized transfer tasks with known psychometric qualities that were not gamified tasks made by the same company that designed the training. However, it should also be noted that any lab-based evaluation would necessarily reduce external validity.

Further, one might also explicitly address the issue that participants of the (unmatched) training and control groups differed in their pre-test performance on the criterion tasks. Although the current sample size is at least moderate, an obvious issue that may arise as a result of the present random group assignment is that random assignment may easily result in unequal performance at pre-test, especially when many tasks are involved (Campbell and Stanley 1966; Green et al. 2014). Future studies could use non-random group assignment protocols specifically designed to reduce imbalance at pre-test (e.g., experimenters create matched pairs based on critical pre-test measures which are then assigned to training and control groups; Spence et al. 2009). From a statistical perspective, there is also the issue of multiple comparison testing since we applied 12 tasks across pre- and post-tests. In order to test our main hypothesis, namely whether there are any significant training effects in the training group versus the control group, we deliberately utilized a MANOVA procedure to rigorously control for alpha cumulation, and additionally ran the same MANOVA without the criterion tasks to rule out that the significant group*time interaction is solely driven by the trained criterion tasks (i.e., to ensure that it indeed represents actual transfer effects). The follow up ANOVA analyses can then be interpreted as an informative guideline to see which of the dependent variables are most sensitive regarding this crucial interaction. We also reasoned that any Bonferroni adjustments (which were designed to correct for statistically independent multiple tests) for these post-hoc ANOVAs would yield too conservative result patterns (with inflated Type II errors) given the high statistical association between dependent variables (see "Results" section).

A final limitation of the present study is related to the interpretation of the size of the observed effects. First, due to the applied nature of our study, we did not utilize standardized working memory tests (with existing test norms) during pre- and post-training assessment. While an analysis based on the difficulty levels achieved for participants indicated a training gain of about two times the size in the training group compared to the control group, it would still clearly be desirable to set up a future large-scale study that includes standardized tests (that ideally selectively index all relevant aspects of cognitive skills) to come up with more interpretable training gain assessments. Second, even if more interpretable data (e.g., with respect to the number of objects that can maximally be held in working memory) exists, it is still an open question to what extent such training gains transfer to real-world skills. While the inclusion of the CFQ in the present study is certainly a first step into the right direction, more research effort is certainly needed to come up with more definite conclusions.

Summary

In general, we assume that the present study is a systematic evaluation of the effectiveness of a specific computerized, home-based brain training under ecologically valid conditions involving tasks that are assumed to tap into working memory capacity and updating. Our results represent evidence against the claim that cognitive training effects are strictly limited to the trained (criterion) tasks (e.g., Owen et al. 2010). Instead, the training appeared to address the underlying cognitive faculties (span/updating in working memory), as evidenced by substantial effects in near-transfer tasks, while effects in far-transfer tasks were not as conclusive as near-transfer effects. Of course, the present study is clearly not able to answer the overarching general question of whether effects of commercial web-based mental trainings have the potential to substantially transfer to daily-life activities. Nevertheless, the data at least suggest that training effects may not be strictly limited to the specific training tasks only, in this sense providing a preliminary response to the statement that "evidence of general and enduring positive effects on the people's minds … has remained elusive" (A consensus on the brain training industry from the scientific community 2014).

Acknowledgements We would like to thank www.neuronation.com and Rouwen Hirth for their technical support during data collection.
Correspondence concerning this article should be addressed to Tilo Strobach, Medical School Hamburg, Department of Psychology, Am Kaiserkai 1, 20457 Hamburg, Germany. E-mail may be sent to tilo.strobach@medicalschool-hamburg.de.

Compliance with Ethical Standards

Conflict of Interest TS and LH have full positions at Medical School Hamburg and at University of Würzburg, respectively. The authors have no financial interest in this study and have no financial disclosure to the participating company. They do not hold shares in this company.

Appendix A

Table 6 List of first appearances of presented TV news issues in Sessions 1–21 of control group treatment (all issues were primarily aired in 1994)

Session	1st issue	2nd issue
1	October 01	September 02
2	August 03	September 03
3	August 04	September 05
4	August 06	August 07
5	September 07	September 08
6	September 09	August 10
7	September 10	August 12
8	August 13	September 13
9	August 14	September 15
10	September 16	August 17
11	September 17	August 18
12	September 18	June 19
13	August 19	August 20
14	August 21	September 22
15	July 24	August 24
16	September 24	September 25
17	August 26	October 01
18	July 27	August 27
19	July 28	August 28
20	September 28	July 29
21	September 30	August 31

Appendix B

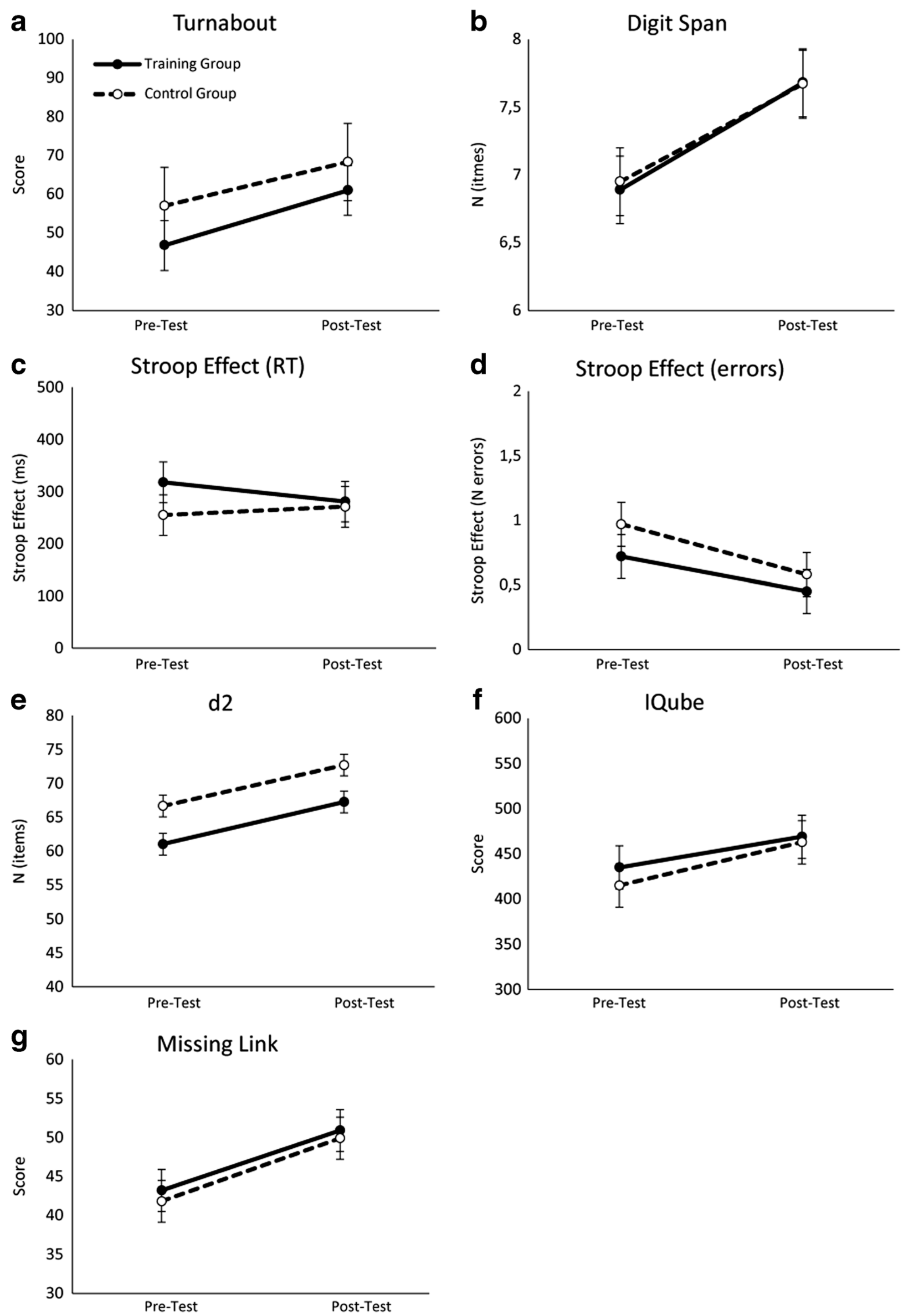


Fig. 4 Overview of the non-significant group*time interactions based on the matched samples data. Performance in the tests **a** Turnabout, **b** Digit span, **c** Stroop Effect (in reaction times [RT]), Stroop Effect (in error rates), **e** d2, **f** IQube, and **g** Missing Link is illustrated for pre-test and post-test sessions as well as for the training and the control groups. Error bars represent SEs

References

A consensus on the brain training industry from the scientific community. (2014). *Retrieved from* http://longevity3.stanford.edu/blog/2014/10/15/the-consensus-on-the-brain-training-industry-from-the-scientific-community-2/.

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., et al. (2013). Video game training enhances cognitive control in older adults. *Nature, 501*(7465), 97–101.

Arnett, J. A., & Labovitz, S. S. (1995). Effect of physical layout in performance of the trail making test. *Psychological Assessment, 7*(2), 220–221.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic Bulletin & Review, 22*(2), 366–377.

Baddeley, A. (1986). *Working memory.* Oxford: Clarendon Press.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), 829–839.

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology, 63*, 1–29.

Blomqvist, N. (1977). On the relation between change and initial value. *Journal of the American Statistical Association, 72*(360a), 746–749.

Boot, W. R., Blakely, D. P., & Simons, D. J. (2011). Do action video games improve perception and cognition? *Frontiers in Psychology, 2*, 226.

Brickenkamp, R. (1962). *Test d2: Aufmerksamkeits-Belastungs-Test*. Göttingen: Hogrefe.

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parkes, K. R. (1982). The cognitive failures questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology, 21*(1), 1–16.

Buckley, D., Codina, C., Bhardwaj, P., & Pascalis, O. (2010). Action video game players and deaf observers have larger Goldmann visual fields. *Vision Research, 50*(5), 548–556.

Campbell, D. T., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological Review, 97*(3), 404–431.

Caruso, J. C. (2004). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *European Journal of Psychological Assessment, 20*(3), 166–171.

Castel, A. D., Pratt, J., & Drummond, E. (2005). The effects of action video game experience on the time course of inhibition of return and the efficiency of visual search. *Acta Psychologica, 119*(2), 217–230.

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review, 17*(2), 193–199.

Cronbach, L. J., & Furby, L. (1970). How we should measure" change": or should we? *Psychological Bulletin, 74*(1), 68–80.

Dahlin, K. I. E. (2013). Working memory training and the effect on mathematical achievement in children with attention deficits and special needs. *Journal of Education and Learning, 2*(1), 118–133.

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., & Nyberg, L. (2008a). Transfer of learning after updating training mediated by the striatum. *Science, 320*(5882), 1510–1512.

Dahlin, E., Nyberg, L., Bäckman, L., & Neely, A. S. (2008b). Plasticity of executive functioning in young and older adults: Immediate training gains, transfer, and long-term maintenance. *Psychology and Aging, 23*(4), 720–730.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450–466.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135–168.

Dye, M. W., Green, C. S., & Bavelier, D. (2009). Increasing speed of processing with action video games. *Current Directions in Psychological Science, 18*(6), 321–326.

Engle, R. W., Carullo, J. J., & Collins, K. W. (1991). Individual differences in working memory for comprehension and following directions. *The Journal of Educational Research, 84*(5), 253–262.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423*(6939), 534–537.

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*(6), 1465–1478.

Green, C. S., & Bavelier, D. (2006b). Enumeration versus multiple object tracking: the case of action video game players. *Cognition, 101*(1), 217–245.

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science, 18*(1), 88–94.

Green, C. S., Strobach, T., & Schubert, T. (2014). On methodological standards in training and transfer experiments. *Psychological Research, 78*(6), 756–772.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: a new hypothesis. *Trends in Cognitive Sciences, 11*(6), 236–242.

Hardy, J. L., Nelson, R. A., Thomason, M. E., Sternberg, D. A., Katovich, K., Farzin, F., & Scanlon, M. (2015). Enhancing cognitive abilities with comprehensive training: a large, online, randomized, active-controlled trial. *PLoS One, 10*(9), e0134467.

Hedge, C., Powell, G., Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*.(In press)

Holmes, J., Gathercole, S. E., Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental science, 12*(4), 9–15.

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*(19), 6829–6833.

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y.-F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning—Implications for training and transfer. *Intelligence, 38*(6), 625–635.

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Developmental Science, 12*(6), 978–990.

Karbach, J., Strobach, T., & Schubert, T. (2015). Adaptive working-memory training benefits reading, but not mathematics in middle childhood. *Child Neuropsychology, 21*(3), 285–301.

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—a review. *Psychological Bulletin, 136*(5), 849–874.

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences, 14*(7), 317–324.

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., et al. (2005). Computerized training of working memory in children with ADHD-a randomized, controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*(2), 177–186.

Koenen, T., Strobach, T., & Karbach, J. (2016). Working memory training. In T. Strobach & K. J. (Eds.), *Cognitive training: An overview of features and applications*. New York: Springer.

Kristjánsson, Á. (2013). The case for causal influences of action videogame play upon vision and attention. *Attention, Perception, & Psychophysics, 75*(4), 667–672.

Li, R., Polat, U., Makous, W., & Bavelier, D. (2009). Enhancing the contrast sensitivity function through action video game training. *Nature Neuroscience, 12*(5), 549–551.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin, 109*(2), 163–203.

Mahncke, H. W., Connor, B. B., Appelman, J., Ahsanuddin, O. N., Hardy, J. L., Wood, R. A., et al. (2006). Memory enhancement in healthy older adults using a brain plasticity-based training program: a randomized, controlled study. *Proceedings of the National Academy of Sciences, 103*(33), 12523–12528.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49*(2), 270–291.

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review, 18*(1), 46–60.

Nouchi, R., Taki, Y., Takeuchi, H., Hashizume, H., Akitsuki, Y., Shigemune, Y., et al. (2012). Brain training game improves executive functions and processing speed in the elderly: a randomized controlled trial. *PLoS One, 7*(1), e29676.

Oberauer, K. (2009). Design for a working memory. *Psychology of Learning and Motivation, 51*, 45–100.

Oldham, P. (1962). A note on the analysis of repeated measurements of the same subjects. *Journal of Chronic Diseases, 15*(10), 969–977.

Olesen, P. J., Westerberg, H., & Klingberg, T. (2004). Increased prefrontal and parietal activity after training of working memory. *Nature Neuroscience, 7*(1), 75–79.

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., et al. (2010). Putting brain training to the test. *Nature, 465*(7299), 775–778.

Rabipour, S., & Raz, A. (2012). Training the brain: fact and fad in cognitive and behavioral remediation. *Brain and Cognition, 79*(2), 159–179.

Rebok, G. W., Carlson, M. C., & Langbaum, J. B. (2007). Training and maintaining memory abilities in healthy older adults: traditional and novel approaches. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 62*(1), 53–61.

Richmond, L. L., Morrison, A. B., Chein, J. M., & Olson, I. R. (2011). Working memory training and transfer in older adults. *Psychology and Aging, 26*(4), 813–822.

Rogosa, D. (1995). Myths and methods: "myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Hillsdale: Lawrence Ealbaum Associates.

Salminen, T., Frensch, P., Strobach, T., & Schubert, T. (2016a). Age-specific differences of dual n-back training. *Aging, Neuropsychology, and Cognition, 23*(1), 18–39.

Salminen, T., Kühn, S., Frensch, P. A., & Schubert, T. (2016b). Transfer after dual n-back training depends on striatal activation change. *Journal of Neuroscience, 36*(39), 10198–10213.

Schmiedek, F. (2016). Methods and designs. In T. Strobach & J. Karbach (Eds.), *Cognitive training: An overview of features and applications*. New York: Springer.

Schmiedek, F., Bauer, C., Lövdén, M., Brose, A., & Lindenberger, U. (2010a). Cognitive enrichment in old age: web-based training programs. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry, 23*(2), 59–67.

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010b). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: findings from the COGITO study. *Frontiers in Aging Neuroscience, 2*, 27.

Schubert, T., Finke, K., Redel, P., Kluckow, S., Müller, H., & Strobach, T. (2015). Video game experience and its influence on visual attention parameters: an investigation using the framework of the theory of visual attention (TVA). *Acta Psychologica, 157*, 200–214.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin, 138*(4), 628–654.

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do "brain-training" programs work? *Psychological Science in the Public Interest, 17*(3), 103–186.

Spence, I., Yu, J. J., Feng, J., & Marshman, J. (2009). Women match men when learning a spatial skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 1097–1103.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading research quarterly 21*(4), 360–407.

Steinborn, M., Langner, R., Flehmig, H. C., & Huestegge, L. (2016). Everyday life cognitive instability predicts simple reaction time variability: analysis of reaction time distributions and delta plots. *Applied Cognitive Psychology, 30*(1), 92–102.

Steinborn, M., Langner, R., Flehmig, H., Huestegge, L. (2017). Methodology of performance scoring in the d2 sustained-attention test: Cumulative-reliability functions and practical guidelines. *Psychological Assessment* (In press).

Strobach, T., & Karbach, J. (2016). *Cognitive training: an overview of features and applications*. New York: Springer.

Strobach, T., Frensch, P. A., & Schubert, T. (2012a). Video game practice optimizes executive control skills in dual-task and task switching situations. *Acta Psychologica, 140*(1), 13–24.

Strobach, T., Frensch, P. A., Soutschek, A., & Schubert, T. (2012b). Investigation on the improvement and transfer of dual-task coordination skills. *Psychological Research, 76*(6), 794–811.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review, 120*(3), 439–471.

Thorndike, E. L. (1924). The influence of the chance imperfections of measures upon the relation of initial score to gain or loss. *Journal of Experimental Psychology, 7*(3), 225–232.

Torous, J., Staples, P., Fenstermacher, E., Dean, J., & Keshavan, M. (2016). Barriers, benefits, and beliefs of brain training smartphone apps: an internet survey of younger US consumers. *Frontiers in Human Neuroscience, 10*, 180.

Tu, Y. K., & Gilthorpe, M. S. (2007). Revisiting the relation between change and initial value: A review and evaluation. *Statistics in Medicine, 26*(2), 443–457.

van Muijden, J., Band, G. P., & Hommel, B. (2012). Online games training aging brains: Limited transfer to cognitive control functions. *Frontiers in Human Neuroscience, 6*, 221.

von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language, 69*(1), 36–58.

Walberg, H. J., & Tsai, S.-L. (1983). Matthew effects in education. *American Educational Research Journal, 20*(3), 359–373.

Wechsler, D. (2008). *Wechsler adult intelligence scale-fourth*. San Antonio: Pearson.

Westerberg, H., Jacobaeus, H., Hirvikoski, T., Clevberger, P., Östensson, M.-L., Bartfai, A., & Klingberg, T. (2007). Computerized working memory training after stroke—a pilot study. *Brain Injury, 21*(1), 21–29.

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory, and how can we measure it? *Frontiers in Psychology, 4*, 433.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education, 15*(1), 345–422.