Psychological Research (2014) 78:756–772
DOI 10.1007/s00426-013-0535-3

REVIEW

On methodological standards in training and transfer experiments

C. Shawn Green · Tilo Strobach · Torsten Schubert

Received: 10 July 2013 / Accepted: 6 December 2013 / Published online: 18 December 2013

Abstract The past two decades have seen a tremendous surge in scientific interest in the extent to which certain types of training—be it aerobic, athletic, musical, video game, or brain trainer—can result in general enhancements in cognitive function. While there are certainly active debates regarding the results in these domains, what is perhaps more pressing is the fact that key aspects of methodology remain unsettled. Here we discuss a few of these areas including expectation effects, test–retest effects, the size of the cognitive test battery, the selection of control groups, group assignment methods, difficulties in comparing results across studies, and in interpreting null results. Specifically, our goal is to highlight points of contention as well as areas where the most commonly utilized methods could be improved upon. Furthermore, because each of the sub-areas above (aerobic training through brain training) share strong similarities in goal, theoretical framework, and experimental approach, we seek to discuss these issues from a general perspective that considers each as members of the same broad "training" domain.

Introduction

For as long as there has been dedicated study of human perception and cognition there has been interest in whether these capabilities can be improved via training (James, 1890; Thorndike & Woodworth, 1901). Over the past decade, this interest has been spurred tremendously by findings suggesting that many obstacles that were previously believed to stand directly in the path of such improvements may in fact be surmountable. For instance, one dominant framework in the field of neural plasticity during the mid- to second-half of the 20th century posited that the brain is capable of large-scale plastic changes only early in life and afterwards becomes somewhat impervious to change (e.g., 'critical' or 'sensitive' periods—Wiesel & Hubel, 1965). Given such a viewpoint, there would be no way to improve basic processing capacities in adulthood or old age through training, as these systems would have become rigid. However, recent work has established that plasticity that had appeared "lost" through maturation, can be at least partially restored via genetic, pharmacological, or even behavioral means (for a review see Bavelier, Levi, Li, Dan, & Hensch, 2010)). Thus, while it is still believed that the brain becomes progressively less malleable over time, it is now clear that even the elderly brain retains sufficient capacity for plasticity to support some degree of improvement.

A second potential roadblock standing in the way of enhanced cognitive or perceptual abilities is the "curse of specificity" (Green & Bavelier, 2008). This refers to the fact that while individuals tend to show improved performance on a task given appropriate training, little to no benefits of this training are seen on new tasks (even if they are seemingly similar to the trained task). Such task specific learning has been shown in nearly all fields of psychology from motor control, to problem solving, reasoning, general cognition, and education (Barnett & Ceci, 2002; Detterman & Sternberg, 1993; Schmidt & Bjork, 1992; Tremblay, Houle, & Ostry, 2008). This type of specificity

C. S. Green (✉)
Department of Psychology, Games+Learning Society, University of Wisconsin-Madison, Madison, WI 53706, USA
e-mail: csgreen2@wisc.edu

T. Strobach · T. Schubert
Department of Psychology, Humboldt University, Berlin, Rudower Chaussee 18, 12489 Berlin, Germany

has been perhaps most intensively investigated though in the domain of perceptual learning. Here, specificity has been observed when learning many low-level features such as orientation, spatial frequency, motion direction, and retinal location (Ball & Sekuler, 1982; Fahle, 2004; Fiorentini & Berardi, 1980).

Yet, recent research suggests that training can indeed produce broad, rather than highly specific, learning. In particular, over the past decade instances of general training effects (i.e., transfer effects) have become pervasive in the literature. For example, there is now a substantial body of evidence demonstrating that aerobic training can improve performance on a wide variety of tasks tapping executive control, spatial abilities, and speed of processing (for a review see Hillman, Erickson, & Kramer, 2008). Similarly, athletic training has been associated with a variety of enhancements in perceptuo-motor skills (for a review see Mann, Williams, Ward, & Janelle, 2007) and some research indicates that musical training increases abilities far beyond music—for instance on measures of fluid intelligence (Schellenberg, 2004). Finally, playing one sub-genre of video games, "action video games," has been shown to lead to improvements in skills ranging from low-level vision (e.g., contrast sensitivity/acuity—Li, Polat, Makous, & Bavelier, 2009) to higher-level executive functions (e.g., task switching/multitasking—Strobach, Frensch, & Schubert, 2012, for a review see Green & Bavelier, 2012).

While many studies in these training domains have been correlational/cross-sectional in nature, critically this literature also includes a large number of experimental studies to establish a causal relationship between the given types of experience and their cognitive outcomes. In fact, the breadth of the effects observed in many of these domains is such that they may be sufficient to be of practical, real-world benefit. Greater aerobic fitness and music education are associated not only with better performance on lab tests, but also with better performance in scholastic settings, e.g., mathematics achievement (Davis et al., 2011; Vaughn, 2000). Finally, action video games have been shown to improve visual performance in individuals with amblyopia (Li, Ngo, Nguyen, & Levi, 2011), surgical ability in endoscopic surgeons (Schlickum, Hedman, Enochsson, Kjellin, & Fellander-Tsai, 2009), and reading performance in children with dyslexia (Franceschini et al., 2013). While there are occasional studies where transfer effects have not been observed (for instance see Blumenthal et al., 1991; Hill, Storandt, & Malley, 1993 for failures to observe effects of aerobic training on cognitive function; Boot, Kramer, Simons, Fabiani, & Gratton, 2008 for a failure to observe effects of action video game training on cognitive function; Neufeld, 1986 for a failure to observe a relationship between music training and mathematics abilities in a sample of kindergarten children), meta-analyses indicate significant effects in each of these domains across tasks and labs (Colcombe & Kramer, 2003; Ferguson, 2007; Vaughn, 2000). These results, and many others, thus provide a reason for optimism that training paradigms can be developed that produce significant enhancements in cognitive function. This emerging belief has resulted in an explosion in the creation and assessment of various dedicated cognitive intervention paradigms designed with the goal of producing more general cognitive enhancements (Bergman Nutley et al., 2011; Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Schmiedek, Lovden, & Lindenberger, 2010; Smith et al. 2009).

Thus, while there is clear potential in training research, many questions still exist. For instance, do individual forms of training produce broad rather than specific behavioral effects (see for instance Klingberg, 2010 and Shipstead, Redick, & Engle, 2012b for differing views on working memory training or Nouchi et al., 2013 and Lorant-Royer, Munch, Mescle, & Lieury, 2010 for conflicting opinions on brain training games such as "Brain Age")? Perhaps more critically though, there remain crucial questions about how to design a study to truly demonstrate that generalization of training effects has or has not been achieved. While issues related to methodology have always had a prominent place in experimental psychology (Campbell & Stanley, 1966; Cook & Campbell, 1979), the methodology employed in fields examining transfer effects has recently been the topic of as much debate as the results of the studies themselves (Boot, Blakely, & Simons, 2011; Schubert & Strobach, 2012; Shipstead, Hicks, & Engle, 2012a; Shipstead et al., 2012b).

Here, we will discuss several of these issues and criticisms, not necessarily in an attempt to describe our view of "proper" methodology, but to suggest that it is unlikely that there is a "one-size"-fits-all methodology that can be utilized. It is instead the case that the proper methods will be closely tied to the research questions being addressed. Furthermore, our explicit goal is to discuss issues that are important across all research domains that study training and transfer effects. Although there exists a number of recent methodological critiques in individual sub-domains (e.g., Shipstead et al., 2012b—working memory training, Boot et al., 2011—action video game training, Rabipour & Raz, 2012—brain training), we think it is valuable to consider these issues from a perspective that encompasses all of these fields as the methodological questions, concerns, and tradeoffs are shared. These include issues related to expectation effects, test–retest effects, control group selection, and subject allocation among others.

Expectation effects

It is often the case that when discussing cognitive training methodology, authors attempt to draw links to medical/pharmacological interventions where the standard method is the double-blind placebo-controlled randomized control trial. Unfortunately, the medical trial analogy breaks down immediately, as one of the primary differences between standard medical trials and cognitive interventions is that in a cognitive intervention, there is no way to "blind" subjects to the content of their training. It is essentially a tautology: subjects that are doing a task know what task they are doing (though, as we discuss below, they may not necessarily know why they are doing the task). Resulting expectation effects (also known as demand characteristics—Orne, 1962)—wherein awareness of a study's hypothesis alters subjects' behavior in the study—create a potential confound in interpreting the results. Several critics (Boot et al., 2011; Boot, Simons, Stothart, & Stutts, 2013; Kristjansson, 2013) have thus argued that effects that have been attributed to differences in experience could in fact be due to subjects being aware that they were expected to have better or worse performance. This expectation can be either a consequence of their previous experience (i.e., action gamers versus non-action gamers, musicians versus non-musicians, athletes versus non-athletes, expert chess players versus non-experts, aerobically fit individuals versus aerobically unfit individuals) or a consequence of their training experience (i.e., in the "brain training group", in the "action video game training group", in the "aerobic activity group", etc.). In both cases, the hypothesis is that subjects alter their behavior to match the expectations created by their group membership.

However, such an expectation-effect based hypothesis requires several conditions to be met. First, subjects need to be aware of the basic hypothesis under consideration. Because subjects in training and transfer experiments are usually not made explicitly aware of the hypotheses prior to or during the experiment, the concern is thus that they will have deduced the correct hypothesis. Furthermore, beyond simple awareness of the basic hypothesis, subjects also need an explicit understanding of how the hypothesis should be expressed in the data (so as to modify their behavior accordingly). In a training and transfer experiment, this further requires that the subject know how they are expected to perform relative to some other group (and therefore they need to have an estimate of how that group 'should' perform as well). For example, action game playing subjects would need to know not only that they were expected to "outperform" non-game players, but also they would need to know that the improvement should be manifested by having larger compatibility effects in visual search displays (Green & Bavelier, 2003), better performance only on dual tasks and not on single tasks (Strobach et al. 2012) or a non-linear enhancement in reaction time coupled with no change in accuracy as compared to non-game playing subjects (Green, Pouget, & Bavelier, 2010). In the case of aerobic activity, the subjects would need to know that improvements should be greater in executive function than in speed of processing (and they would therefore need to know which interaction terms in which tasks are indicative of speed of processing versus executive functioning (Colcombe & Kramer, 2003). It is thus unclear whether subjects would come to the lab with beliefs related to the specific hypotheses under consideration (rather than incredibly broad beliefs—e.g., in Boot et al., 2013a) or know the ways in which these hypotheses should be born out in data.

In fact, even those experts nominally in the field are not always good at making correct predictions. For example, when discussing expectation effects, Boot, Blakely and Simons (2011) described a hypothetical training experiment involving an action video game trained group and a Tetris (control) trained group: "Following training, participants view (but do not perform) two transfer tasks: (a) a fast-paced task in which participants detect targets flashed in the visual periphery (the useful field of view, or UFOV) and (b) a task in which participants mentally rotate block-like shapes. The Tetris training group likely would predict that their training would improve their mental rotation performance and the action game training group likely would predict better UFOV performance…" (Boot et al., 2011). This example is of particular interest because a study nearly identical to that described above had actually already been done by Feng Spence and Pratt (2007), with the hypothesis and the result that action gaming improves both UFOV performance and mental rotation performance. Because this finding of an effect on both tasks contradicts the hypothesis of Boot et al., it clearly shows that even individuals with interest in the field can make errors regarding actual experimental hypotheses.

Beyond knowing the experimental hypothesis and the way hypotheses would be born out in data, it must be possible for participants to alter their performance so as to match the subjects' desires. While the studies in the expectation effects literature ensure this is the case (e.g., a study may be one of visual preference where subjects are free to select an image on the left or the right of the screen on each trial as in (Nichols & Maner, 2008), it is less clear that this is possible on the "performance" tasks. If, for instance, it is possible to improve measured working memory capacity via desire to do so alone, it suggests significant issues for the general working memory literature. Indeed, if one's working memory capacity could be substantially altered by one's beliefs about what one's working memory capacity should be, assessments of

working memory capacity would be reasonably meaningless as it would be unclear whether these assessments reflected the capacity of subjects who "wished" to have high capacities or who did not "wish" to have high capacities (see also Schubert & Strobach, 2012 for a discussion of this issue). Similarly, it would be somewhat problematic if the reason chess experts demonstrate expert performance in chess is in fact not due to knowledge structures acquired via deliberate practice (Ericsson, Krampe, & Tesch-Romer, 1993), but is instead due to their belief that they should be good at chess.

While the overall issue of expectation effects is absolutely worth considering, the available empirical findings uniformly point against a major role for expectation effects. First, if expectancy effects played a significant role in the generation of transfer effects after action video game training, one would expect such effects in every study on this issue. However, the action gaming literature includes several examples wherein action video game players (who had been overtly selected) consistently show no benefits—for instance, on tasks that measure the dynamics of certain aspects of attention (Castel, Pratt, & Drummond, 2005; Hubert-Wallander, Green, Sugarman, & Bavelier, 2011). Furthermore, many different studies on expert versus non-expert gamers (Clark, Fleck, & Mitroff, 2011; Colzato, van den Wildenberg, Zmigrod, & Hommel, 2013; Donohue, Woldorff, & Mitroff, 2010; Dye & Bavelier, 2010; Dye, Green, & Bavelier, 2009; Trick, Jaspers-Fayer, & Sethi (2005)) have been performed with completely covert recruitment (i.e., the subjects' game playing habits were assessed unrelated to testing and thus could not have contaminated their test performance). The results have been typically consistent with results from studies with non-covert recruiting methods. In Trick et al. (2005) the subjects did not fill out gaming questionnaires (this was done by their parents), and yet the same action video game advantages on the multiple-object tracking task were observed that were later found in Green and Bavelier (2006), a study that used overt recruitment. Similarly, both Dye and Bavelier (2010) and Dye et al. (2009) utilized blind recruitment in their child samples and showed the same effects of action video game experience as in studies in which either the recruitment was not blind and/or the questionnaires on game playing habits were finished prior to testing. In fact, it is notable that in these two latter studies, the behavioral performance in the 14–17-year-old age group was nearly identical to the 18–22-year-old age group, despite the fact that the former was recruited blindly while the latter was selected for video game playing status. Such a pattern of results is inconsistent with what would be anticipated if expectation effects were a serious concern.

Finally, it is also unclear how expectancy could produce certain neuroplastic changes that have been observed either in expert gamers or after action video game training. For instance, Bavelier, Achtman, Mani and Focker (2011) showed that increasing task difficulty (in a visual search task) resulted in increasing activity in the fronto-parietal network of brain areas in non-action game playing individuals (i.e., greater task difficult = greater engagement of the attentional system). However, the same increase in task difficulty led to almost no change in the fronto-parietal network of expert action video game players. This pattern of results is consistent with the proposal that expert action video game players benefit from a more efficient attentional system and is inconsistent with the expectation hypothesis [simply "trying harder" would almost certainly lead to substantial activation of the fronto-parietal network; see also (Krishnan, Kang, Sperling, & Srinivasan, 2013; Mishra, Zinni, Bavelier, & Hillyard, 2011) for examples of functional differences associated with action game expertise]. Furthermore, in a video game training study, Wu et al. (2012) showed that individuals who were trained and who improved at a first-person shooter video game, demonstrated clear changes in late visual area ERP components.

Despite the fact though that the available literature suggests that expectation effects may not be a likely confound in interpreting existing results in training studies, it is nevertheless the case that attempting to measure their influence would be useful for the field. Unfortunately, how exactly to measure the effects such knowledge may have on performance is not currently clear. For example, one substantial issue complicating the enterprise is that subjects are not always truthful when they are probed for such information. For instance, in a study by Nichols and Maner (2008) each participant in the study ($N = 100$) was deliberately and explicitly made aware of the experimental hypothesis by a confederate prior to testing. This was accomplished by having each subject sit in a waiting room before entering the testing room. While they were waiting, the "previous subject" (actually a confederate) emerged and told the waiting subject that the experimental hypothesis was that people would tend to disproportionately like images that are presented to their left-hand side. While subjects' behavior in the test was by and large consistent with an attempt to confirm the hypothesis suggested by the confederate, zero out of the one hundred participants admitted to knowing about the hypothesis when they were probed at the end of the experiment. Thus, while those in the training and transfer domain can certainly administer such suspicion probes, it is uncertain whether they will truly control for the effect in question. Even if subjects report that they did not infer the hypothesis, a critic could still argue that the suspicion probe simply did not uncover the true state of affairs. Furthermore, one possibility that is often overlooked is that

suspicion probes themselves can be subject to expectation effects and thus the truthfulness of "yes" responses can be met with similar apprehension as are "no" responses. One could easily imagine a subject reasoning that: "If an experimenter is asking me if I thought I knew the hypothesis, they are clearly hoping I'll say something, so I'll come up with something on the spot despite the fact that I had no expectations during the task". The same problematic issue holds for recent attempts of Boot et al. (2013a, b) who asked subjects for their expectations after they played different games. There is no way to know whether these are thoughts the subjects would have had without being asked.

What then are possible suggestions for modifications to the methods that have been most commonly employed in the literature to date? One possibility would be to create a control group only for expectation effects (i.e., wherein the experimenter or another confederate deliberately and explicitly tells subjects in this control group that they are actually in the "active" condition where benefits are expected). This still leaves the possibility though that subjects will not believe this information. Another possible suggestion is to always include tasks wherein performance is known to be modifiable by changes in expectation, but that the treatment would not be expected to modify. For instance, using the example task above (i.e., Nichols & Maner, 2008), one could imagine having a confederate inform subjects in an aerobic training study that more aerobically fit individuals tend to prefer images presented on the left side of the screen. If such a pattern of results were indeed observed (which would not otherwise be an expected effect of aerobic training), it would suggest that such expectations could be a contaminant in other results as well. However, it is essential to note that such a pattern would not necessarily imply that other results were similarly contaminated, nor would the failure to find such a pattern necessarily imply that other results were not contaminated.

Thus, utilizing a mixture of approaches may be the most viable course because each method of either controlling for, or eliciting proof of, expectation effects has serious limitations. However, the potential benefits will need to be weighed against the costs (as some of these methods could greatly increase expenditures both in money and in time) and may only be necessary in more mature fields (i.e., there is little virtue in controlling for possible confounding causes of group differences until one knows whether such differences are observed as a function of a new training paradigm at all).

Test–retest effects

Several authors (Boot et al., 2011; Kristjansson, 2013) have recently suggested that in order for a study to be valid, both the treatment and the control group must show significant test–retest improvements. In this vein, Boot et al. (2011) remark that according to "learning theory," "participants typically improve when performing a cognitive task for a second time" (pp. 3). Similarly Kristjansson (2013) remarks that "Visual and attentional performance usually improves with practice…". For these authors, in cases where the treatment group shows a significant improvement from pre- to post-test, but the control group does not, the difference between the groups may not be due to transfer from some active treatment to the testing, but may instead be the result of some mechanism wherein the control training prevents test–retest improvements (i.e., the active treatment has no effect on improvement, but the control training instead simply blocks learning). Unfortunately, these statements are simply inconsistent, both with the available literature on perceptual and cognitive learning and with the methodology that provides the best opportunity to observe transfer.

To address the question—"are control subjects performing 'as expected' when they do not demonstrate significant test–retest improvements"—we must first understand the exact nature of the criticism. Although this is not explicitly stated, based upon the totality of the author's arguments, it must be the case that they expect significant test–retest effects in essentially each and every case (Boot et al., 2011; Kristjansson, 2013). Otherwise there would be no cause for criticism, as the existing literature is consistent with the belief that improvement "often" occurs. Examining this stronger hypothesis then, should we expect significant test–retest effects in every situation independent of context? No, we should not. And indeed, many papers in the field of perceptual learning clearly demonstrate that one single exposure to a task is often insufficient to drive significant improvements on the task. For example, in a dot motion direction discrimination learning study by Ball and Sekuler (1982), subjects were trained to determine whether two sequentially presented clips of moving dot stimuli moved in the same or different directions. Subjects first pre-tested on 8 directions of motion (spaced evenly across 360°). They were then trained on just one direction. After three sessions of training on this one direction, they were once again tested on all 8 directions. Subjects showed a significant increase in performance on the trained direction, but no change on the untrained directions despite the fact that they were "performing the test for the second time" (interestingly the experiment then continued on through several more cycles of training on one direction and testing on the others, with the same lack of retest advantage seen throughout). Similarly, in a recent elegant demonstration of training conditions that produce transferable effects (using a 'double training' paradigm), Xiao et al. (2008) first showed that

performing a simple contrast discrimination task twice leads to absolutely no observable benefits in performance. In fact, one could argue that the entire literature on specificity of learning (in perceptual learning and beyond) is built upon the foundation that a single pre-test session is not typically sufficient to drive significant improvements on a single post-test session (Redick et al., 2013; Watkins & Smith, 2013).

Furthermore, beyond the literature on transfer of learning there is a reasonably extensive literature demonstrating failures to observe learning even with substantial practice. For instance, in an influential set of studies by Herzog and Fahle (1997), not only was one session with feedback insufficient to observe an increase in vernier acuity performance, no learning was seen after 640 trials without feedback (see also Seitz, Nanez, Holloway, Tsushima, & Watanabe, 2006).

Thus, subjects simply are not expected to improve on every task after a single exposure independent of the specifics of the task. The failure to observe significant test–retest effects is often the pattern one would expect from subjects performing "as expected" and thus critiques should be wary of simplifying this complex question to an extent to which it becomes thoroughly inconsistent with existing knowledge. It is beyond the scope of this review to list the situations where learning is not expected, but factors such as the task domain, presence or absence of feedback, the difficulty of the task, and the time between the first and second experiences with the task will matter greatly when predicting whether test–retest effects are "expected."

We now turn to the question of whether test–retest effects are something to aspire to. While this appears to be the belief of some authors (Boot et al., 2011; Kristjansson, 2013), we take the opposite view: Test–retest effects are an explicit enemy of those who seek to observe transfer effects and, therefore, steps should be taken to minimize these effects. To clarify this point, consider the following hypothetical example (see Fig. 1). Subjects are pre-tested on some Task A, which involves the cognitive process—α. The treatment group then trains on an experimental task that involves the cognitive process of interest (α) as well as a number of unrelated processes (γ, δ, ε, and Ψ). The other half of the subjects (the control group) train on a control task that involves only the unrelated processes (γ, δ, ε, and Ψ). After training, all subjects are post-tested on Task A. As most views on transfer suggest that the degree of transfer is a function of the similarity in processing demands between the training and the transfer test (going back to Thorndike & Woodworth, 1901 or see Singley & Anderson, 1989; Taatgen, 2013 for a more modern view), the obvious hypothesis is that transfer effects should be observed in the treatment group (which shares process α with the test), while no transfer effects should be observed in the control group (which shares no processes in common with the test). As is clear in Fig. 1 though, the size of the transfer effect depends strongly on the extent to which there is learning on the test itself. Large amounts of learning on the pre-test means that there is little left to learn through training and consequently there will be nearly no difference between the treatment and control groups at post-test (Fig. 1a, b). Conversely, when steps are taken to reduce learning at pre-test, the learning that occurs during training can be observed as transfer at post-test (Fig. 1c, d). Given this conceptualization, larger test–retest effects will always predict relatively smaller observed transfer effects. Essentially by demanding significant test–retest effects, both Boot et al. (2011) and Kristjansson (2013) are insisting on reducing the probability of observing transfer.

So what does the framework above suggest about design in a training and transfer experiment? The simplest principle is that conditions that produce test–retest effects such as using many trials, providing informative feedback, and having a short period of time between test and retest, should be avoided. It is perhaps not surprising that one failed training experiment in the action video game training literature (Boot et al., 2008) is one in which there were significant testing effects (to the extent that both the treatment and control groups at the final test outperformed expert action game players). These principles also inexorably argue against attempts to estimate dose–response curves via repeated testing (i.e., test → train → retest → train → retest → etc.). While one cycle of test–retest may not always lead to significant learning, more and more testing increases the probability of learning on the test itself and, therefore, diminishes the probability of observing transfer effects (with again, the severity of the concern depending on the tendency of the particular task to produce learning). The only way to avoid these issues when attempting to estimate dose response curves is to perform studies such as those undertaken by Jaeggi et al. (2008), wherein different groups of subjects are trained for different lengths of time.

Finally, the framework above also suggests a different way of analyzing data than what is most commonly employed in the field. In typical training and transfer experiments we compare performance averaged across all pre-test trials with performance averaged across all post-test trials. By employing such an analysis, we are making the implicit assumption that performance is stable across the entire pre-test and the entire post-test. However, it is quite possible that we are often averaging over a learning curve. This makes assessing transfer problematic (particularly because, given the framework above, subjects who start worse at the post-test should show more improvement,

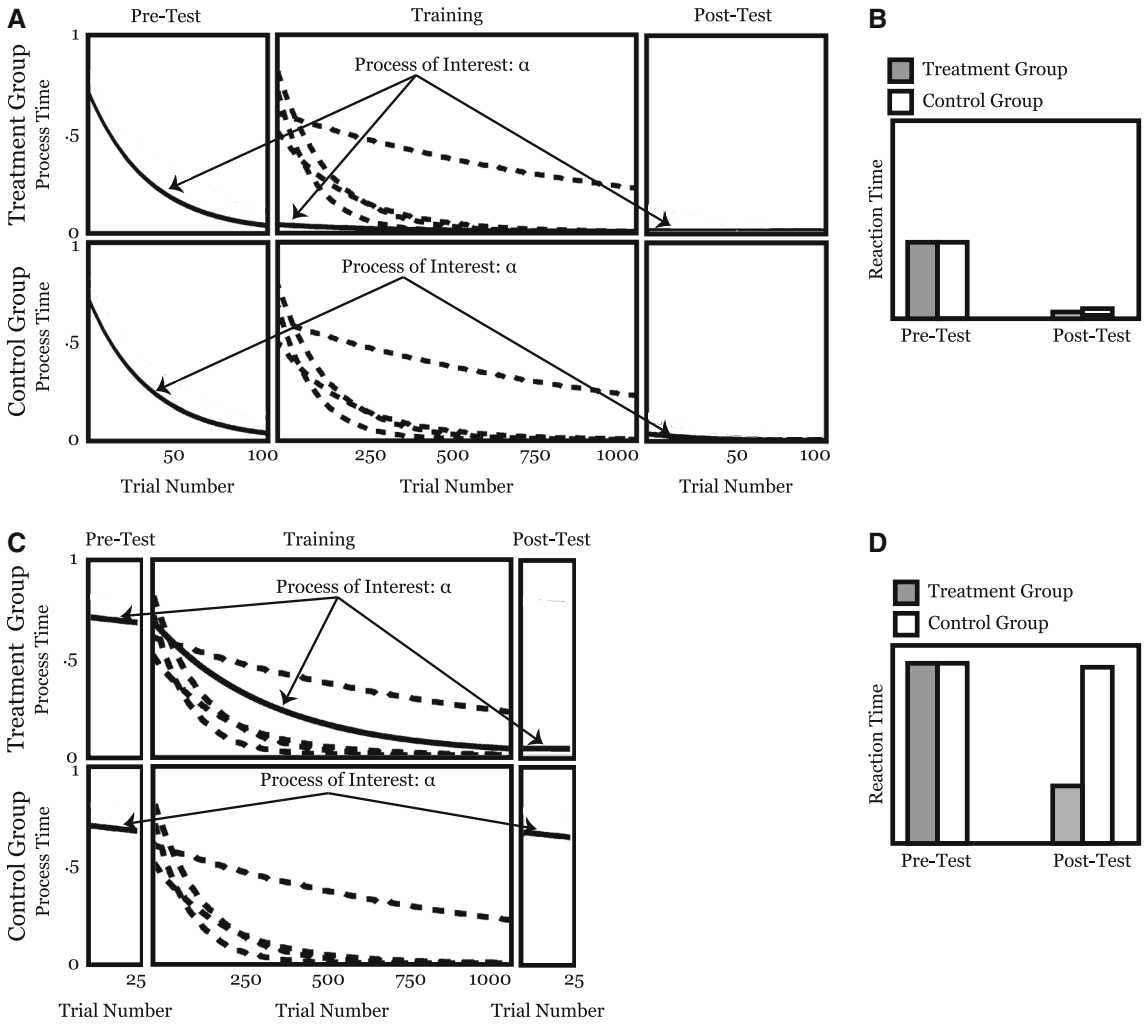

Fig. 1 Effect of test–retest improvements on ability to observe transfer: in this hypothetical example, the question at hand is whether some form of treatment training will result in significant transfer/benefits on another test. This hypothetical test involves a single cognitive process α (*solid black line*). Transfer is expected from the treatment training to the test because the experimental treatment also involves this process (α), while the control does not. **a** *Left panel* here subjects learn a significant amount during the pre-test (i.e., their reaction times improve significantly). *Middle panel* during training, while there is significant learning on the unrelated processes (*dashed lines*), there is very little possible improvement left for process α, even though the training does tax this process. *Right panel* neither group has much room left for improvement at post-test. **b** When performance is averaged over the pre-test and post-test and the groups are compared, significant test–retest effects are clear, but there is no differential effect of training. **c** *Left panel* here, testing has been altered to minimize learning. No feedback is given to reduce the rate of learning and testing is much shorter (25 trials versus 100 trials). *Middle panel* there is thus still a significant amount left to learn about process α for the treatment group. *Right panel* following training, the treatment group has finished learning about process α, while the control group still has room for improvement. **d** When performance is averaged, there is a small, but non-significant improvement in the control group, while the treatment group shows a significant drop in RT at post-test. Note: the learning in all cases is exponential as per (Dosher & Lu, 2007)

as there is more to learn). Instead, a more theoretically appropriate approach would be to fit learning curves to pre- and post-test data (preferably at the level of individual subjects). In this way, it would be possible to ask whether performance on the first trial of the post-test is consistent with what would be expected given performance on the last trial of the pre-test (see for example Jeter, Dosher, Petrov, & Lu, 2009, where full learning curves are fit both to the learning and the transfer tests).

Size of the test battery

Another significant point of contention in all training and transfer fields is the size of the test battery, with several recent papers (Melby-Lervag & Hulme, 2013; Shipstead et al., 2012b) criticizing the field for utilizing test batteries that are too small (often only a single test). While these papers have very eloquently laid out the case for a larger test battery, the basic arguments are worth reiterating here.

The most critical virtue of a large test battery is that it allows inferences to be made at the level of processes or constructs rather than at the level of individual tasks (Engle, Tuholski, Laughlin, & Conway, 1999; von Bastian & Oberauer, 2013). For instance, if one wishes to demonstrate an increase in "working memory capacity," it is not sufficient to show enhancements on a single task that, at best, partially loads on working memory; there are simply too many reasons unrelated to working memory that might cause an improvement on that one single task. However, if after training subjects show improvements on many different working memory tasks (especially if they do so in proportion to the extent to which the tasks are known to load on working memory—Colom, Martinez-Molina, Shih, & Santacreu, 2010), this would constitute much stronger evidence that the training did in fact enhance the working memory. Furthermore, a large test battery also allows the insertion of tests of "no interest" (i.e., tasks that are not expected to be affected by the training regimen). While researchers typically only include tests of abilities that they expect to change as a function of their treatment, tests of no interest have the potential to be incredibly informative. For instance, consider the assessment of a paradigm designed with the goal of improving working memory. If subjects are tested on multiple tests of working memory as well as multiple tests of an ability unrelated to working memory (e.g., visual search) before and after training, there are two possible patterns of data that could emerge at post-testing (assuming the treatment provides any benefit at all as compared to an active control).

Pattern 1

The treatment group improved by a greater amount than the control on the tests of working memory, while no differences between groups were seen on the tests of visual search.

Pattern 2

The treatment group improved by a greater amount than the control on the tests of working memory AND the tests of visual search.

In both cases, there is a significant effect of the treatment on working memory performance. However, the most likely mechanism underlying this outcome is clearly quite different given Pattern 1 versus Pattern 2.

While it is important to highlight the need to utilize an appropriately large test battery, it is likewise vital to recognize that there are potential downsides to an ever-larger test battery (beyond the simple fact that an increasing number of measures—whether they be related or not—increases the probability of Type I errors). It is not necessarily the case that if three working memory tasks are better than one, that five is better than three, ten better than five, and so forth. What are the potential drawbacks? The first possible issue is fundamentally the same as that discussed in the section above on test–retest effects: more testing will tend to lead to more learning on the transfer tests, which will in turn reduce the potential to observe transfer from the treatment. The only difference here is that learning in this case would be hierarchical in nature and occurs at the level of the individual tasks (Ahissar, Nahum, Nelken, & Hochstein, 2009; Bavelier, Green, Pouget, & Schrater, 2012). Indeed, it is necessarily the case that tests that measure similar constructs share structure at some level of abstraction. Thus, by experiencing many of these tests, learning could occur at these hierarchical levels. For instance, non-verbal intelligence tests (which are nearly always pattern discovery tests) tend to have strong similarity in the types of patterns that are present (e.g., add a component across columns, subtract a component across rows, reorder a component in a systematic way across rows, etc.). It is, therefore, possible that individuals who take many IQ tests will become highly familiar with the set of patterns that are often present and they will thus have less opportunity to demonstrate a benefit of training per se.

A second possible issue is related to cognitive depletion (Baumeister, Bratslavsky, Muraven, & Tice, 1998; Muraven & Baumeister, 2000; Schmeichel, 2007) or cognitive fatigue type effects (Bryant & Deluca, 2004). Indeed, there is convincing evidence that performance on tasks, particularly those that involve control, inhibition, and other executive type functions (i.e., exactly those functions that are of most interest to the field), diminishes as a function of time spent on these tasks (Salminen, Strobach, & Schubert, 2012). For example, Holtzer, Shuman, Mahoney, Lipton and Verghese (2011) showed that cognitive fatigue was associated specifically with reduced performance on the executive control component of the Attentional Network Test. In large test batteries that take several hours, these effects can therefore be of substantial concern, particularly given that cognitive fatigue creates potential confounds in both directions. If one does observe significant transfer effects in a treatment group, but not in a control group, these could be accounted for by training that does not improve the construct of interest, but instead merely improves the ability to resist cognitive fatigue. Conversely, cognitive fatigue could disproportionately affect better subjects (i.e., those that would otherwise show transfer—as would be predicted if fatigue leads to a proportional rather than an additive reduction in performance) and thus eliminate the ability to observe transfer that would otherwise be present.

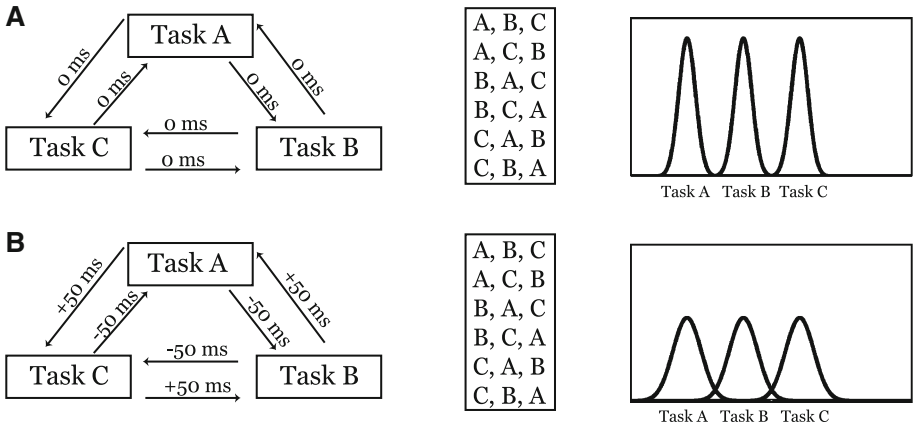

Fig. 2 Effect of temporal order effects on estimate of inter-subject variability: **a** *left panel* here there is no effect of order on task performance. *Middle panel* subjects are nonetheless run in a pseudorandom order. *Right panel* estimates of performance and inter-subject variability on the three tasks can be acquired (Task A faster than Task B faster than Task C, with equivalent inter-subject variability). **b** Here the tasks interact with one another. For instance, having just performed Task A leads one to be 50 ms faster on Task B, while having just performed Task B leads one to be 50 ms slower on Task A (note: order effects are symmetrical here for convenience, although they may not be in practice). *Middle panel* subjects are again run on one of the six possible orders. *Right panel* the counterbalancing has removed the bias (i.e., the estimates of the means are equivalent to that found in Fig. 3a). However, estimates of inter-subject variability are much higher (because the statistical model is not taking into account the order effects and instead attributing these to intrinsic variability). While the order effects could theoretically be removed (e.g., modeled as an interaction), the number of subjects necessary for such an analysis to be feasible is quite large (likely unfeasible) with only three tasks, let alone with a more extensive battery

A third possible issue is that performance may not be independent of the order of previous tests (Klauer & Mierke, 2005). In a large test battery, we are not measuring performance on Test A, Test B, Test C and so forth. Instead, we are measuring performance on Test B given that the subject has already taken Test A, performance on Test C given that the subject has already undergone Test A and B, and so forth. This leads us to temporal dependence effects—sometimes also known as carryover effects or serial order effects (Brooks, 2012). These effects can take any number of possible forms—some which will facilitate performance and some which will inhibit performance. For example, similarity or differences in response structure or the relationship between stimuli and responses can lead to both positive and negative effects on new tasks (Osgood, 1949). Similarly, contrast effects—wherein a task seems easier or more difficult by comparison with a previous task—could result in subjects being more or less confident in their current task performance and thus fundamentally alter their performance on the new task (Plous, 1993).

All of these dependence effects though, share the general property that they create variability in subject performance that is unaccounted for and, thus, may reduce the ability to observe a significant transfer effect. Although authors will often attempt to counter-balance the presentation order of tasks to reduce biases associated with temporal dependence effects, two key points are worth noting. First, while counterbalancing will potentially reduce the overall influence of bias by allowing one to average over the various task sequence effects (e.g., facilitations and inhibitions), which eventually come to a null, it does not eliminate the presence of the dependence effects and thus will continue to represent variability that is incorrectly attributed to individual differences between subjects in statistical models (see Fig. 2). Second, counterbalancing becomes increasingly impractical as the size of a battery grows (i.e., with a battery of three tests there are only six combinations that need to be counterbalanced, but with a battery of eight tests there are already over 40,000 combinations).

Lastly, as is true of cognitive depletion effects, the possibility of temporal dependence effects adds another potential confound in the interpretation of any improvements that are observed. Namely, any improvements in performance may be a result of an enhancement in the ability to prevent the detrimental influence of previous tasks on current task performance rather than enhancements in the base construct of interest. That is the subjects may not have truly "improved" on any of the cognitive tasks, but may have instead reduced the extent to which they allow previous tasks to negatively impact their current behavior; that the latter is indeed conceivable has been shown by many studies on task switching, (e.g., Kray & Lindenberger, 2000).

What do these issues suggest for experimental design? First, in terms of selecting the size of the test battery the goal should be to utilize a battery that is "just the right size." That size depends critically on the questions at hand,

the extent to which the various tests load on a factor of interest, etc. If one's goal is to test possible effects at the construct level, then there is a minimum number of tasks that are necessary to allow for such an inference to be made. Conversely, if one has a tightly defined research question regarding a single process (e.g., in the case of dual-task coordination skills, see Liepelt, Strobach, Frensch, & Schubert, 2011; Strobach, Frensch, Soutschek, & Schubert, 2012), then it will be sufficient to use a smaller number of tasks, which are most closely related to the process in question and which are most accepted as paradigmatic. By and large, under these preconditions, one should use the smallest (and not largest) possible number of tests, which is necessary to obtain the current research goals.

Finally, although some confounds, such as hierarchical learning, are somewhat unresolvable (though manipulations such as altering the background context can make it more difficult for subjects to learn that information is organized hierarchically), confounds such as depletion, cognitive fatigue, and carry over effects can be minimized by spreading out testing rather than utilizing single long testing sessions. Care can also be taken to spread out tests of similar constructs that might lead to greater fatigue or carry over effects (as was done in Redick et al., 2013).

Control groups

The selection of the proper control group is one of the most fundamental issues in training and one of the most contentious. There is general agreement that active control groups are necessary, as simple test–retest/no contact/passive control groups fail to rule out too many possible confounds to allow results to be meaningfully interpreted. The types of active control groups employed though are almost always sub-domain or even investigator specific. For instance, in the sub-domain of music training, the control is often another activity from the arts, such as drama training (Schellenberg, 2004). In the sub-domain of aerobic training, the control is often something else physical in nature such as toning/stretching (Voss et al., 2010). In the action video game literature, the control is typically another video game from a non-action genre (such as simulation games like The Sims or puzzle games like Tetris). For "brain training," tasks often do double-duty as treatment and control—where one training task is predicted to improve one area of processing such as speed while another is predicted to improve a different area such as reasoning (Smith et al. 2009).

At the conceptual level, the "right" control is one that incorporates all of the possible causes of an improvement that are of "no interest". However, there is not currently complete agreement as to what mechanisms are of "no interest" and this is one reason why there is no agreement as to the "right" control group. For instance, it is clearly the case that improvements due to test–retest effects are of no interest, and thus any control group must take the tests the same number of times and in the same manner as the treatment group. And, as stated above, most believe that improvements due to motivation, arousal, etc. are serious confounds when trying to evaluate the efficiency of an intervention, and thus any control group should be 'active' (as opposed to no contact, for a discussion see Green & Bavelier, 2012). However, there is debate over what type of "active" control is most appropriate. Some groups strongly believe that a proper control group must have an adaptive level of difficulty (i.e., increase in difficulty as the subject improves at the task). For instance, Redick et al. (2013) utilized an adaptive visual search task as their control-training task, Jaeggi, Buschkuehl, Jonides and Shah (2011) utilized an adaptive knowledge/quiz-type paradigm, and the control games utilized in the action gaming literature (e.g., Tetris) naturally become more difficult as the player performs better (Green & Bavelier, 2003; Strobach et al. 2012). Furthermore, controls in areas like music/aerobic activity/video games are typically adaptive by nature (i.e., drama lessons tend to increase in difficulty according to good pedagogy, toning exercises increase in demand with improvements in muscle function). However, other groups take a different tactic and use a non-adaptive version that is otherwise identical to the treatment task—for instance, a treatment group that performs an N-back task where the level of "N" increases with increasing performance/ability and a control group that performs the exact same task except that "N" remains at a low-level throughout training (Brehmer, Westerberg, & Backman, 2012; Klingberg et al., 2005). This latter tactic has the advantage of very closely matching many components of the trained task (stimulus characteristics/material, motor responses, task instructions, etc.), but it is unclear whether it can control for other items of "no interest" such as motivation and/or arousal.

What then are possible guidelines for selecting control paradigms? First, we would argue that the selection of the control paradigm/groups should depend on the specific research aim of the study. Thus, if one is testing a product, i.e., a "brain trainer" designed to be commercially sold as something that enhances cognition, then the proper control is very different than if the goal is to uncover underlying mechanisms of transfer. For a product, the proper control should be analogous to the medical field where the standard is not typically a totally inactive placebo control, but is instead the best current treatment (i.e., when hospitals are testing a new cancer drug they do not compare a new drug to a sugar pill, they compare the new drug to the current standard of care). So for those who wish to sell a product to

the public, it is essential to say more than "our program is better than something that controls for the bare minimum of confounds" (or worse that "our program is better than doing nothing at all"). Instead, the goal should be to say that, "our program is better than what is currently (and in many cases freely) available"—be that action video games, aerobic training, musical training, the programs used in the ACTIVE/COGITO trials (Schmiedek et al., 2010; Smith et al. 2009), etc.

For those doing basic research, although it would be incredibly useful to have a "standard" control that all labs interested in transfer could utilize, it is unlikely that there will ever be such a one-size-fits-all approach. If, in our lab, we are interested in testing the hypothesis that large amounts of spatial and temporal uncertainty must be present in video games to observe transfer to new visuo-spatial tasks, we cannot use the same control as a group of researchers who are interested in testing the hypothesis that long-term training on a set of reasoning heuristics will promote improvements in reasoning on new problem sets. Similarly, the ideal control in a study utilizing a population of institutionalized elderly individuals will not be the same as the ideal control in a study utilizing a population of college-aged adults.

Finally, it is important when designing a control task to control for the most likely confounds, but also to create a control task that is distinct enough from the experimental task to maximize observable training effects. This follows the principle of effect maximization in experimental psychology (Huber, 2009). Thus, our goal should be to find a balance between manipulating as few processes of interest as is possible, while still utilizing experimental conditions that would allow for the detection of a difference in the parameter of interest. If we created a treatment and a control condition that, according to a sophisticated theoretical model, differ only in one fine-grained cognitive construct (e.g., treatment uses a_{i}, while the control uses a_{ii}), this design would have the potential to tell us something very meaningful about this particular cognitive construct. However, our instantiations of these constructs might not in fact be large enough to create sufficiently different data sets to actually observe significant differences (particularly given the relatively small samples utilized in most long-lasting training research). Therefore, when creating the control conditions we ought to maximize the probability of observing an effect by creating clear differences in the process of interest (i.e., not testing a_{i} against a_{ii}, but instead comparing a_{i} with a_{iiii}—or to use a more tangible example, not testing a treatment with large working memory demands against a control with moderate working memory demands, but instead against a control with minimal working memory demands).

Random or non-random group assignment

One of the first tenets students learn in every introductory methods class is that random sampling and random group assignment is the "gold standard" of experimental methods. However, we would argue that using purely random group assignment, data in those fields concerned with training and transfer often becomes impossible to interpret. The most obvious issue that arises as a result of purely random group assignment is that random assignment may easily result in unequal performance at pre-test, as exemplarily illustrated in Fig. 3. The probability of unequal performance increases when the sample size is relatively small, as is very often the case in the domains of training and transfer research (Campbell & Stanley, 1966). These pre-test differences (even if they do not rise to the level of significantly different) in turn bring a multitude of possible confounds into play. For instance, if the treatment group's performance at pre-test is better than the control group's performance (Fig. 3 right, 1st panel), a 'successful' intervention (i.e., the treatment group improves by more than the control group) could be attributed to the fact that control subjects often learn very little from tasks that far exceed their abilities. Conversely, if the intervention is 'unsuccessful' (i.e., the treatment group does not improve by more than the control group; not illustrated), this could be attributed to ceiling effects (or the fact that the treatment started on the shallow part of an exponential learning curve). Indeed, the treatment effect may need to be quite large to overwhelm this type of selection-regression effect (Campbell & Stanley, 1966). Arguably though, the more troubling case is that in which the treatment group's performance is initially worse than the control group's performance (Fig. 3 right, 2nd panel). When this is the case, the predicted effects will result in a crossover interaction, which makes the inference process extremely difficult. Can one truly construe this as being indicative of a "successful" intervention? Is the proper interpretation instead that the effects represent simple regression to the mean? With purely post hoc analyses there is no effective way to perfectly unpack these possibilities.

Fortunately, there are non-random methods of group assignment specifically designed to reduce imbalance at pre-test (see Fig. 3 right, 3rd and 4th panel). Probably the most commonly used methods historically speaking are "blocking" and "pairing" (Addelman, 1969; Feldt, 1958). In a randomized blocked-design, subjects are first divided into reasonably homogenous subgroups. These subgroups are then randomly split and assigned to the different conditions. For instance, if, on the dependent variable of interest at pre-test, subject scores ranged from 0 to 200, purely random assignment could create quite unequal groups (see Fig. 3). In a randomized block design, subjects would first be stratified based upon their pre-test scores

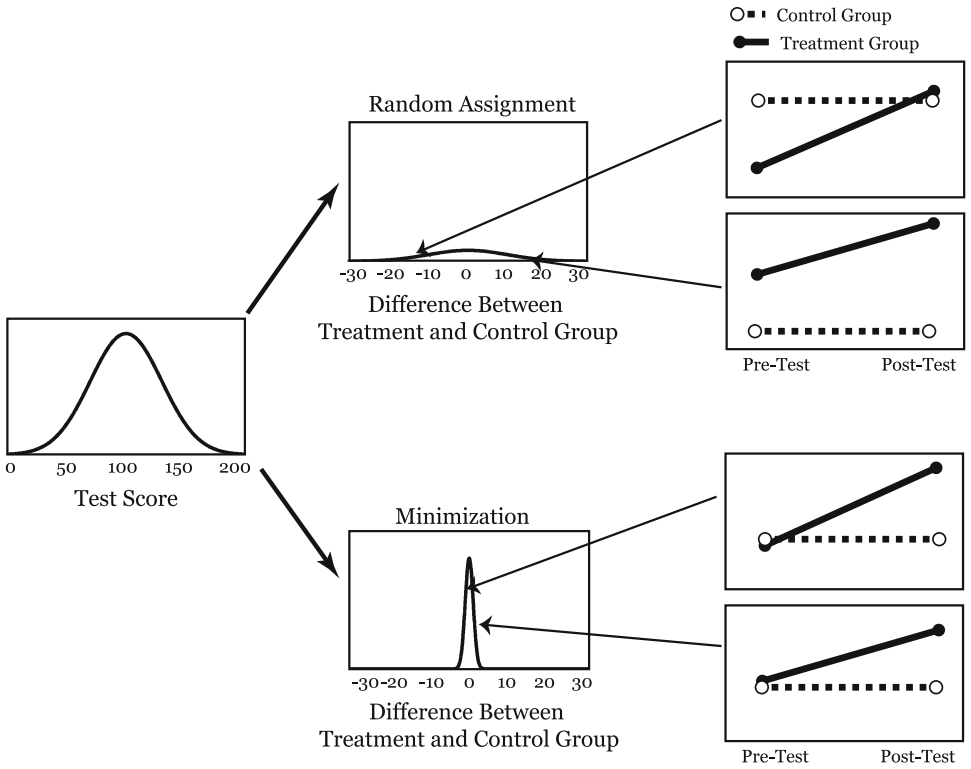

Fig. 3 Randomization versus minimization in group assignment: from the *left*: subject test scores have a mean of 100 and a standard deviation of 30. *Middle top* when subjects from this population are randomly assigned to a control and an active treatment group (groups of 16 subjects), on average there is no difference between the groups, however, individual samples may have large differences from zero (i.e., the treatment group may vastly outperform the control group or vice versa). *Middle bottom*—conversely, using a minimization algorithm reduces the chance of a large difference existing between the two groups. *Top right* large differences between groups at pre-test lead to difficulties in interpreting eventual results. If the treatment group is worse than the control group at pre-test, then the expected pattern of results is a "catching up" effect or a "crossover" effect (leading to concerns about regression to the mean). If the treatment group is better than the control group at pre-test, then one concern is that the test is simply too difficult for the control group to show any improvement. *Bottom right* minimization avoids these issues (i.e., the two graphs are nearly identical—in one case the control group is only slightly better than the treatment at pre-test, while in the other, the treatment group is only slightly better than the control group)

(i.e., all those who scored from 0 to 10 are put in one subgroup, all of those who scored from 11 to 20 are put into a second subgroup and so on), before being randomly split and assigned to a condition. A paired design is conceptually the same; the only difference is that rather than creating subgroups—the experimenter creates matched pairs. In the training and transfer literature, Spence, Yu, Feng and Marshman (2009) used this type of design to examine the effect of action video game training on spatial cognition in males as compared to females. In their study, a cohort of males and females were first pre-tested on a measure of spatial cognition. Male/female pairs with closely matched pre-test scores were then invited to continue into the training phase thus ensuring roughly equivalent pre-test scores across the two sexes (see also Redick et al., 2013 or Loosli, Buschkuehl, Perrig, & Jaeggi, 2012 for additional examples).

Beyond these more classic approaches, a method known as minimization (Pocock & Simon, 1975; Taves, 1974) has gained some prevalence in clinical trials and could be of use in the training and transfer domain. The general principle underlying minimization is that while the first subject is assigned to a group at random, subsequent subjects are assigned so as to reduce imbalance between the groups (Chen & Lee, 2011; Smith, 1984). Under this general category label, there are many specific algorithms (Saghaei, 2011). There are methods that can be utilized for cases where all prognostic factors are known for all subjects (i.e., all pre-test scores, etc.) prior to group assignment, or in cases where enrollment is rolling. In the most straightforward instantiation, subjects are simply added to the group that minimizes the mean difference between the groups. From there though, the algorithms can greatly increase in complexity to, for instance, minimizing differences in multiple variables (potentially weighted by their reliability) or to matching both means and standard deviations of the groups (as unequal variance between groups violates the assumptions of most parametric statistical tests). In addition, this minimization method can lead to different sample sizes.

It is worth noting though that none of these approaches are without issue (see also Campbell & Stanley, 1966). For instance, blocking approaches typically require that all subjects be pre-tested prior to any subject beginning training (i.e., the researcher needs to know the pre-test scores for all subjects before the subgroup divisions can be formed). This can be difficult for labs that are not equipped to train the entire experimental cohort simultaneously. It is also the case that there will be diminishing returns associated with adding increasing numbers of blocking variables as well as the potential for increased cost. For pairing approaches, a certain number of subjects that are pre-tested will not be utilized because no matched pair can be formed (e.g., in Spence et al., 2009, 43 subjects were pre-tested to find 10 matched pairs). This can represent a significant cost in time and effort. Finally, all of the approaches come with associated potential pitfalls in analysis that need to be carefully considered (Kahan & Morris, 2012a, b; Zhao, Hill, & Palesch, 2012). One remaining question for the field going forward is whether it should become standard to perform additional parallelization between groups on measures beyond the specific tests of interest. For instance, one could imagine circumstances where it would be useful to match groups according to IQ, working memory capacity, or basic learning ability (e.g., if the goal is to test the effect of a treatment on problem solving).

Difficulty in making inferences across studies

One of the major issues in the field is that the results of studies testing the effect of one training regimen do not allow for inferences to be made regarding training regimens not under consideration. For instance, the success or failure of a music training paradigm cannot be used to alter current estimates of the probability that a working memory training paradigm will be successful. More unfortunately still, we are not only lacking in the ability to connect across sub-domains, to some extent we cannot even draw inferences within a sub-domain. For instance, the success or failure of one "brain training" paradigm cannot speak to the likelihood of success or failure of any other regimen that falls under this label. As an example, the results of the highly cited brain training paper by Owen et al. (2010), led many to the conclusion that "brain-training games don't work" (Rutherford, 2010—although it is important to note that Owen et al. were much more careful with their wording). What can actually be concluded from this study is that the methodology employed by the authors (e.g., "brain training games" that were 'gamified' versions of standard psychology paradigms and performed training at home) resulted in no statistically significant improvements on a small battery of tests of reasoning, short-term memory, etc. From this data, it is not possible to infer what would have occurred had a slightly different population been employed or if more than just a few hours of training were required, let alone to extrapolate to the entirety of "computerized brain trainers." A similar argument could be made for the recent paper by (Redick et al., 2013), which cannot truly be taken as a failure to replicate the work by (Jaeggi et al., 2008). Given the number of changes made to the regimen (e.g., the nature and the extent of the test battery), this must be considered as a separate result all together (whether one or the other is more indicative of the 'ground truth' will be for future research to determine).

Therefore, although the current push in the field is overwhelmingly toward translational research (whether to combat age-related cognitive decline or to reduce symptoms associated with attentional disorders), the current state of affairs calls much more strongly for work at the theoretical end. We assume that, as a field, we are in a similar position as was the field of gene therapy in its earliest stages. By the 1970s techniques existed to insert segments of foreign DNA into mammalian cells and thus to permanently alter their function. However, while it was evident that there was the potential for the development of an incredibly powerful tool to fight human disease (Friedmann & Roblin, 1972), it was similarly clear that the knowledge necessary to fulfill this tool's vast potential was lacking (of the genome in general, of the relationship between genes, the proteins they code for, and the eventual disease state that they mediate, etc.). And indeed, rather than haphazardly and indiscriminately entering into translational gene therapy trials, nearly two decades of foundational knowledge was accrued before the first human gene marking study was approved (and only in the past 5–10 years has the technique been used to great success (Sheridan, 2011). In the area of cognitive enhancement, it is not remotely evident that we understand the mechanisms well enough to wield tools at our disposal in practical settings although we have a significant number of them.

What then are some options going forward? The most obvious answer is that a significantly greater amount of data needs to be collected. Only with a large amount of data on a wide variety of different paradigms (treating the issues associated with non-significant results appropriately—see the next section) can we begin to understand the space of factors inherent in cognitive training paradigms and then in turn to discover the parts of the space that produce the most effective training regimens.

Acceptance of the null hypothesis

A final issue concerns the observation of null results and how they should be interpreted. One might have imagined

that when a null result is found in a field still in its infancy (as is true of the cognitive enhancement field) the standard approach would be to look for factors that may have obscured the finding of a possible training effect (i.e., to ensure that the null result was not due to an idiosyncrasy of a particular design choice). Yet, in current discussions we often find the opposite approach—a push toward constraining designs so as to minimize the probability of finding an effect (many of which are discussed earlier in the paper). Part of this trend has undoubtedly arisen in response to the proliferation of "brain training" products now being sold and advertised with only the flimsiest (if any) evidence backing their claims. However, it is unclear whether this trend is truly in the best interests of the overarching field, particularly given the fact that most studies are designed in such a way that only a positive outcome will actually be informative. That is, studies are often significantly underpowered and thus null results provide essentially no useful information—this should not be taken to suggest that null results should not be published, only that studies should be designed so that null results are worth publishing.

Take for instance the case of training time. If, in a hypothetical study, 5 h of aerobic training results in greater changes in executive function than 5 h of stretching, this is informative. However, if, given the same methodology, no significant differences were found between the groups, this is far less revealing. This latter pattern of results certainly cannot be used to infer that "aerobic training" has no effect on executive function. It would thus be useful if studies were designed to ensure that 'insufficient length of training' could not be the cause of a null result (perhaps by calculating when a plateau has been reached in learning on the training task—although even then further learning improvements may occur after a plateau has been reached).

The same logic can be applied to factors discussed earlier in the manuscript. While there is a certain virtue in designing paradigms to ensure that confounding factors are not the root cause of a positive training effect, it is equally critical to ensure that confounding factors are not the root cause of null training effects.

Conclusions

The past two decades have witnessed a significant increase in the study of training regimens—from aerobic exercise, to musical training, to athletics, to video games, to working memory training, to dedicated 'brain trainers'—that may produce general enhancements in cognitive functioning. Although these regimes are often treated as very separate domains, the questions that have arisen related to methodology ("How can we most convincingly demonstrate that a given regimen has a given effect?") are strongly shared across all of these sub-fields. Some issues that have been raised, such as expectation effects, do not appear to be a significant concern based upon the currently available data, though attempts can nonetheless be made to more fully control for such possibilities. What forms those controls should take however, remains an open question (e.g., it is unclear that simple suspicion probes truly represent value added). Other issues, such as the proper size of the test battery, while quite legitimate, are unlikely to have a one-size-fits-all solution. Large test batteries have both clear virtues (e.g., the ability to make inferences at the level of cognitive constructs rather than at the level of individual tests) as well as potentially large drawbacks (e.g., reductions in power due to order effects or cognitive fatigue) and thus the proper test battery size clearly depends on the goals of the experiment. Similarly, the selection of a control task also depends deeply on the hypotheses under consideration. Finally, there are some areas, such as group assignment, in which the field will continue to improve.

References

Addelman, S. (1969). The generalized randomized block design. *The American Statistician, 23*(4), 35–36.

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society B, 364*, 285–299.

Ball, K. K., & Sekuler, R. (1982). A specific and enduring improvement in visual motion discrimination. *Science, 218*, 697–698.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology, 74*(5), 1252–1265.

Bavelier, D., Achtman, R. L., Mani, M., & Focker, J. (2011). Neural bases of selective attention in action video game players. *Vision Research,*. doi:10.1016/j.visres.2011.08.007.

Bavelier, D., Green, C. S., Pouget, A., & Schrater, P. (2012). Brain plasticity through the life span: Learning to learn and action video games. *Annual Review of Neuroscience, 35*, 391–416.

Bavelier, D., Levi, D. M., Li, R. W., Dan, Y., & Hensch, T. K. (2010). Removing brakes on adult brain plasticity: From molecular to behavioral interventions. *Journal of Neuroscience, 30*(45), 14964–14971.

Bergman Nutley, S., Soderqvist, S., Bryde, S., Thorell, L. B., Humphreys, K., & Klingberg, T. (2011). Gains in fluid intelligence after training non-verbal reasoning in 4-year-old children: A controlled, randomized study. *Developmental Science, 14*(3), 591–601.

Blumenthal, J. A., Emery, C. F., Madden, D. J., Schniebolk, S., Walsh-Riddle, M., George, L. K., et al. (1991). Long-term effects of exercise on psychological functioning in older men and women. *Journal of Gerontology: Psychological Sciences, 46*, 352–361.

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

Boot, W. R., Champion, M., Blakely, D. P., Wright, T., Souders, D. J., & Charness, N. (2013a). Video games as a means to reduce age-related cognitive decline: Attitudes, compliance, and effectiveness. *Frontiers in Psychology, 4*, 31. doi:10.3389/fpsyg.2013.00031.

Boot, W. R., Kramer, A. F., Simons, D. J., Fabiani, M., & Gratton, G. (2008). The effects of video game playing on attention, memory, and executive control. *Acta Psychologica, 129*, 387–398.

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013b). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science, 8*(4), 445–454.

Brehmer, Y., Westerberg, H., & Backman, L. (2012). Working-memory training in younger and older adults: Training gains, transfer, and maintenance. *Frontiers in Human Neuroscience, 6*, 63. doi:10.3389/fnhum.2012.00063.

Brooks, J. L. (2012). Counterbalancing for serial order carryover effects in experimental condition orders. [Research Support, Non-U.S. Gov't]. *Psychological Methods, 17*(4), 600–614. doi:10.1037/a0029310.

Bryant, D. C. N., & Deluca, J. (2004). Objective measurement of cognitive fatigue in multiple sclerosis. *Rehabilitation Psychology, 49*(2), 114–122.

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

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 (Amst), 119*(2), 217–230.

Chen, L. H., & Lee, W. C. (2011). Two-way minimization: A novel treatment allocation method for small trials. *PLoS One, 6*(12), e28604.

Clark, K., Fleck, M. S., & Mitroff, S. R. (2011). Enhanced change detection performance reveals improved strategy use in avid action video game players. *Acta Psychologica (Amst), 136*(1), 67–72. doi:10.1016/j.actpsy.2010.10.003.

Colcombe, S., & Kramer, A. F. (2003). Fitness effects on the cognitive function of older adults: A meta-analytic study. *Psychological Science, 14*(2), 125–130.

Colom, R., Martinez-Molina, A., Shih, P., & Santacreu, J. (2010). Intelligence, working memory, and multitasking performance. *Intelligence, 38*, 543–551.

Colzato, L. S., van den Wildenberg, W. P. M., Zmigrod, S., & Hommel, B. (2013). Action video gaming and cognitive control: Playing first person shooter games is associated with improvement in working memory, but not action inhibition. *Psychological Research, 77*, 234–239.

Cook, T., & Campbell, D. T. (1979). *Quasi-experimental design*. Chicago: Rand McNally.

Davis, C. L., Tomporowski, P. D., McDowell, J. E., Austin, B. P., Miller, P. H., Yanasak, N. E., et al. (2011). Exercise improves executive function and achievement and alters brain activation in overweight children: A randomized, controlled trial. *Health Psychology, 30*(1), 91–98.

Detterman, D. K., & Sternberg, R. J. (1993). *Transfer on trial: Intelligence, cognition, and instruction*. Norwood: Ablex Publishing Corporation.

Donohue, S. E., Woldorff, M. G., & Mitroff, S. R. (2010). Video game players show more precise multisensory temporal processing abilities. *Attention, Perception and Psychophysics, 72*(4), 1120–1129.

Dosher, B. A., & Lu, Z. (2007). The functional form of performance improvements in perceptual learning: Learning rates and transfer. *Psychological Science, 18*(6), 531–539.

Dye, M. W. G., & Bavelier, D. (2010). Differential development of visual attention skills in school-age children. *Vision Research, 50*(4), 452–459.

Dye, M. W. G., Green, C. S., & Bavelier, D. (2009). The development of attention skills in action video game players. *Neuropsychologia, 47*, 1780–1789.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309–331.

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

Fahle, M. (2004). Perceptual learning: A case for early selection. *Journal of Vision, 4*(10), 879–890.

Feldt, L. S. A. (1958). A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika, 23*, 335–353.

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science, 18*(10), 850–855.

Ferguson, C. J. (2007). The good, the bad, and the ugly: A meta-analytic review of positive and negative effects of violent video games. *The Psychiatric Quarterly, 78*(4), 309–316.

Fiorentini, A., & Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency. *Nature, 287*, 43–44.

Franceschini, S., Gori, S., Ruffino, M., Viola, S., Molteni, M., & Facoetti, A. (2013). Action video games make dyslexic children read better. *Current Biology, 23*(6), 462–466.

Friedmann, T., & Roblin, R. (1972). Gene therapy for human genetic disease? *Science, 175*(4025), 949–955.

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

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

Green, C. S., & Bavelier, D. (2008). Exercising your brain: A review of human brain plasticity and training-induced learning. *Psychology and Aging, 23*(4), 692–701.

Green, C. S., & Bavelier, D. (2012). Learning, attentional control and action video games. *Current Biology, 22*, R197–R206.

Green, C. S., Pouget, A., & Bavelier, D. (2010). Improved probabilistic inference as a general mechanism for learning with action video games. *Current Biology, 23*, 1573–1579.

Herzog, M. H., & Fahle, M. (1997). The role of feedback in learning a vernier discrimination task. *Vision Research, 37*, 2133–2141.

Hill, R. D., Storandt, M., & Malley, M. (1993). The impact of long-term exercise training on psychological function in older adults. *Journal of Gerontology, 48*, 12–17.

Hillman, C. H., Erickson, K. I., & Kramer, A. F. (2008). Be smart, exercise your heart: Exercise effects on brain and cognition. *Nature Reviews Neuroscience, 9*, 58–65.

Holtzer, R., Shuman, M., Mahoney, J. R., Lipton, R., & Verghese, J. (2011). Cognitive fatigue defined in the context of attention networks. *Neuropsychology, Development, and Cognition: Section B, Aging, Neuropsychology and Cognition, 18*(1), 108–128. doi:10.1080/13825585.2010.517826.

Huber, O. (2009). *The psychological experiment: An introduction (in German)*. Bern: Hans Huber.

Hubert-Wallander, B., Green, C. S., Sugarman, M., & Bavelier, D. (2011). Changes in search rate but not in the dynamics of exogenous attention in action videogame players. *Attention, Perception, and Psychophysics, 73*(8), 2399–2412.

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., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences, 108*, 10081–10086.

James, W. (1890). *The principles of psychology* (Vol. I). New York: Dover Publications Inc.

Jeter, P. E., Dosher, B. A., Petrov, A., & Lu, Z. L. (2009). Task precision at transfer determines specificity of perceptual learning. *Journal of Vision, 9*(3), 1–13.

Kahan, B. C., & Morris, T. P. (2012a). Improper analysis of trials randomised using stratified blocks or minimisation. *Statistics in Medicine, 31*(4), 328–340. doi:10.1002/sim.4431.

Kahan, B. C., & Morris, T. P. (2012b). Reporting and analysis of trials using stratified randomisation in leading medical journals: Review and reanalysis. [Research Support, Non-U.S. Gov't Review]. *BMJ, 345*, e5840. 10.1136/bmj.e5840.

Klauer, K. C., & Mierke, J. (2005). Task-set inertia, attitude accessibility, and compatibility-order effects: New evidence for a task-set switching account of the implicit association test effect. *Personality and Social Psychology Bulletin, 31*(2), 208–217. doi:10.1177/0146167204271416.

Klingberg, T. (2010). Training and plasticity of working memory [Review]. *Trends in Cognitive Sciences, 14*(7), 317–324. doi:10.1016/j.tics.2010.05.002.

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

Kray, J., & Lindenberger, U. (2000). Adult age differences in task switching. *Psychology and Aging, 15*, 126–147.

Krishnan, L., Kang, A., Sperling, G., & Srinivasan, R. (2013). Neural strategies for selective attention distinguish fast-action video game players. *Brain Topography, 26*(1), 83–97. doi:10.1007/s10548-012-0232-3.

Kristjansson, A. (2013). The case for causal influences of action video game play upon vision and attention. *Attention, Perception, and Psychophysics, 75*(4), 667–672.

Li, R. W., Ngo, C., Nguyen, J., & Levi, D. M. (2011). Video-game play induces plasticity in the visual system of adults with amblyopia. *PLoS Biology, 9*(8), e1001135.

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.

Liepelt, R., Strobach, T., Frensch, P. A., & Schubert, T. (2011). Improved intertask coordination after extensive dual-task practice. *The Quarterly Journal of Experimental Psychology, 64*(7), 1251–1272.

Loosli, S. V., Buschkuehl, M., Perrig, W. J., & Jaeggi, S. M. (2012). Working memory training improves reading processes in typically developing children. *Child Neuropsychology, 18*(1), 62–78.

Lorant-Royer, S., Munch, C., Mescle, H., & Lieury, A. (2010). Kawashima vs "Super Mario"! Should a game be serious in order to stimulate cognitive aptitudes? *European Review of Applied Psychology, 60*(4), 221–232.

Mann, D. T., Williams, A. M., Ward, P., & Janelle, C. M. (2007). Perceptual-cognitive expertise in sport: A meta-analysis. *Journal of Sport and Exercise Psychology, 29*(4), 457–478.

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

Mishra, J., Zinni, M., Bavelier, D., & Hillyard, S. A. (2011). Neural basis of superior performance of action videogame players in an attention-demanding task. *Journal of Neuroscience, 31*(3), 992–998.

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle. *Psychological Bulletin, 126*(2), 247–259.

Neufeld, K. A. (1986). Understanding of selected pre-number concepts: Relationships to a formal music program. *Alberta Journal of Educational Research, 32*(2), 132–139.

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology, 135*(2), 151–165.

Nouchi, R., Taki, Y., Takeuchi, H., Hashizume, H., Nozawa, T., Kambara, T., et al. (2013). Brain training game boosts executive functions, working memory and processing speed in the young adults: A randomized controlled trial. *PLoS ONE, 8*(2), e55518. doi:10.1371/journal.pone.0055518.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776–783.

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review, 56*(3), 132–143.

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.

Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill Education.

Pocock, S. J., & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics, 31*, 103–115.

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

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., et al. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General, 142*(2), 359–379.

Rutherford, A. (2010). Brain-training games don't work. *The Guardian*. Retrieved from http://www.theguardian.com

Saghaei, M. (2011). An overview of randomization and minimization programs for randomized clinical trials. *Journal of Medical Signals and Sensors, 1*(1), 55–61.

Salminen, T., Strobach, T., & Schubert, T. (2012). On the impacts of working memory training on executive functioning. *Front Hum Neurosci, 6*, 166. doi:10.3389/fnhum.2012.00166.

Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychological Science, 15*(8), 511–514.

Schlickum, M. K., Hedman, L., Enochsson, L., Kjellin, A., & Fellander-Tsai, L. (2009). Systematic video game training in surgical novices improves performance in virtual reality endoscopic surgical simulators: A prospective randomized study. *World Journal of Surgery, 33*(11), 2360–2367.

Schmeichel, B. J. (2007). Attention control, memory updating, and emotion regulation temporarily reduce the capacity for executive control. *Journal of Experimental Psychology: General, 136*(2), 241–255. doi:10.1037/0096-3445.136.2.241.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207–217.

Schmiedek, F., Lovden, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience, 2*. doi:10.3389/fnagi.2010.00027.

Schubert, T., & Strobach, T. (2012). Video game experience and optimized executive control skills—On false positives and false

negatives: Reply to Boot and Simons (2012). *Acta Psychologica, 141*(2), 278–280.

Seitz, A. R., Nanez, J. E, Sr, Holloway, S., Tsushima, Y., & Watanabe, T. (2006). Two cases requiring external reinforcement in perceptual learning. *Journal of Vision, 6*(9), 966–973.

Sheridan, C. (2011). Gene therapy finds its niche. *Nature Biotechnology, 29*(2), 121–128. doi:10.1038/nbt.1769.

Shipstead, Z., Hicks, K. L., & Engle, R. W. (2012a). Cogmed working memory training: Does the evidence support the claims? *Journal of Applied Research in Memory and Cognition, 1*, 185–193.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012b). Is working memory training effective? *Psychological Bulletin, 138*(4), 623–654.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge: Harvard University Press.

Smith, R. L. (1984). Sequential treatment allocation using biased coin designs. *Journal of the Royal Statistical Society: Series B, 46*, 519–543.

Smith, G. E., Housen, P., Yaffe, K., Ruff, R., Kennison, R. F., Mahncke, H. W., et al. (2009). A cognitive training program based on principles of brain plasticity: Results from Improvement in Memory with Plasticity-based Adaptive Cognitive Training (IMPACT) study. *Journal of the American Geriatrics Society, 57*(4), 594–603.

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.

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.

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

Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics, 15*, 443–453.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review, 8*, 247–261.

Tremblay, S., Houle, G., & Ostry, D. J. (2008). Specificity of speech motor learning. *Journal of Neuroscience, 28*(10), 2426–2434.

Trick, L. M., Jaspers-Fayer, F., & Sethi, N. (2005). Multiple-object tracking in children: The "Catch the Spies" task. *Cognitive Development, 20*(3), 373–387.

Vaughn, K. (2000). Music and mathematics: Modest support for the oft-claimed relationship. *Journal of Aesthetic Education, 34*(3/4), 149–166.

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

Voss, M. W., Prakash, R. S., Erickson, K. I., Basak, C., Chaddock, L., Kim, J. S., & Kramer, A. F. (2010). Plasticity of brain networks in a randomized intervention trial of exercise training in older adults. Frontiers in Aging Neuroscience, *2*. doi:10.3389/fnagi.2010.00032.

Watkins, M. W., & Smith, L. G. (2013). Long-term stability of the Wechsler intelligence scale for children-fourth edition. *Psychological Assessment,*. doi:10.1037/a0031653.

Wiesel, T. N., & Hubel, D. H. (1965). Comparison of the effect of unilateral and bilateral eye closure on cortical unit responses in kittens. *Journal of Neurophysiology, 26*, 1003–1017.

Wu, S., Cheng, C. K., Feng, J., D'Angelo, L., Alain, C., & Spence, I. (2012). Playing a first-person shooter video game induces neuroplastic change. *Journal of Cognitive Neuroscience, 24*(6), 1286–1293.

Xiao, L., Zhang, J., Wang, R., Klein, S. A., Levi, D. M., & Yu, C. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology, 18*, 1922–1926.

Zhao, W., Hill, M. D., & Palesch, Y. (2012). Minimal sufficient balance—A new strategy to balance baseline covariates and preserve randomness of treatment allocation. *Statistical Methods in Medical Research,*. doi:10.1177/0962280212436447.