Assessing the applied benefits of perceptual training: Lessons from studies of training working-memory

Nori Jacoby
ELSC Center for Brain Sciences, The Hebrew University, Jerusalem
Massachusetts Institute of Technology,
Department of Brain and Cognitive Sciences,
Cambridge, MA, USA

Merav Ahissar
ELSC Center for Brain Sciences and the Department of Psychology,
The Hebrew University, Jerusalem

In the 1980s to 1990s, studies of perceptual learning focused on the specificity of training to basic visual attributes such as retinal position and orientation. These studies were considered scientifically innovative since they suggested the existence of plasticity in the early stimulus-specific sensory cortex. Twenty years later, perceptual training has gradually shifted to potential applications, and research tends to be devoted to showing transfer. In this paper we analyze two key methodological issues related to the interpretation of transfer. The first has to do with the absence of a control group or the sole use of a test–retest group in traditional perceptual training studies. The second deals with claims of transfer based on the correlation between improvement on the trained and transfer tasks. We analyze examples from the general intelligence literature dealing with the impact on general intelligence of training on a working memory task. The re-analyses show that the reports of a significantly larger transfer of the trained group over the test–retest group fail to replicate when transfer is compared to an actively trained group. Furthermore, the correlations reported in this literature between gains on the trained and transfer tasks can be replicated even when no transfer is assumed.

Introduction

In the 1980s to 1990s, many studies reported that visual training effects are specific to the position of the trained stimuli in the visual field (e.g., Ahissar & Hochstein, 1997; Fiorentini & Berardi, 1980; Herzog & Fahle, 1997; Karni & Sagi, 1991) and to the trained orientations (Ahissar & Hochstein, 1993; Fiorentini & Berardi, 1981; Levi & Polat, 1996). These findings were striking from the perspective of the specificity of early sensory cortices, which represent these parameters in a segregated manner. Additionally, training was specific to the trained task (e.g., Ahissar & Hochstein, 1993; Herzog & Fahle, 1997).

However, with the growing awareness of the potential clinical applications of perceptual training (boosted largely by Tallal et al., 1996), scientists studying cognition became increasingly interested in the generalization of training. The rationale was as follows: If we can pinpoint the limitations on an individual's perceptual mechanism, and given that the brain is plastic, we can train this individual on a task directed toward this mechanism and thus improve its performance. This training process was expected to "release the bottleneck" that limits perceptual performance in a broad range of conditions. However, this reasoning turned out to be inconsistent with the findings of learning specificity, as well as with theoretical accounts that viewed task-specific improvement as reflecting a greater ability to decipher task-specific signal from noise (e.g., Ahissar & Hochstein, 2004; Ahissar, Nahum, Nelken, & Hochstein, 2009; Dosher & Lu, 1998).

When viewed from a broader perspective, the perceptual learning literature of the 1980s to 1990s that reported this stimulus and task specificity clearly coincides with the vast earlier literature on skill acquisition in general. This literature consistently showed that acquired expertise is specific to both the trained stimuli and their trained behavioral relevance. One of the best studied cases is that of expert chess players who have vastly enhanced working memory (WM) abilities. However, their enhanced WM skills are specific to chess moves, and do not extend even to

Citation: Jacoby, N., & Ahissar, M. (2015). Assessing the applied benefits of perceptual training: Lessons from studies of training working-memory. *Journal of Vision*, *15*(10):6, 1–13, doi:10.1167/15.10.6.

doi: 10.1167/15.10.6 Received March 16, 2015; published August 25, 2015 ISSN 1534-7362

random arrangements of the same pieces on the chess board (Chase & Simon, 1973).

Perhaps the most desired cognitive skill is intelligence. Intelligence is often considered to be reflected in the common statistical factor underlying performance of academically related reasoning tasks (defined as *g* and often dubbed fluid intelligence, which refers to potential abilities rather than acquired, crystalized skills; Cattell, 1987). This general ability is tightly correlated with WM skills (Kyllonen & Christal, 1990). One of the most extensively studied WM tasks is the *N*-back task, in which individuals are exposed to a sequence of stimuli and are asked to indicate whenever a stimulus is a repetition of a stimulus presented *N* steps backwards. This task is challenging, since any stimulus can be a repetition of the previous one. It thus requires continuous tracking of external stimuli and online updating of the internal representation of the anticipated stimulus. Performance was shown to be highly correlated with performance on intelligence measures such as matrix completion, and is known to have a high load on g (e.g., Conway, Kane, & Engle, 2003; Engle, Kane, & Tuholski, 1999a; Kyllonen & Christal, 1990). This robust correlation led to the hope that intensive training would boost performance on the *N*-back task, which would generalize to enhanced intelligence scores.

The idea that intelligence can be boosted by training on a specific task seems radical since fluid intelligence is thought of as a genetically related trait, although perhaps affected by long-term educational and cultural factors. However, it is not conceptually different from boosting basic perception by enhancing attentional allocation abilities, or improving the efficiency of implicit statistical inferences, which are the outcomes attributed to playing action video games (Green, Pouget, & Bavelier, 2010). The Flynn effect (Flynn, 1987) suggests that fluid intelligence is indeed susceptible to change. The Flynn effect is the observation that performance of tasks such as matrix completion has improved by 2–10 points every decade since the earliest measurements in the 1930s. This improvement is attributed to the fact that education has gradually placed more emphasis on abstract (rather than only concrete), context-free analogies. The Flynn effect indicates some generalization, since individuals are not specifically trained on the matrix completion task, and yet their performance on this task improved. However, WM training studies make their claim to generalization based on a short training period with a very limited set of tasks.

In this paper we analyze two important methodological difficulties undermining studies that have made this claim. The first is what constitutes a good control, and the second is whether correlations between gains on the trained and untrained task provide evidence for transfer. Both questions are highly pertinent to studies on perceptual training, as described below.

Forming and testing a valid control group

Lessons from WM training

In the WM training literature, many studies only include a passive control group in addition to the experimental group trained on some form of the *N*-back task. Passive control refers to a group administered the same series of pre- and posttests, with a similar time interval, but with no intervening training procedure. This group controls for test and retest effects, but does not control for very broad placebo effects.

For example, one of the most frequently cited papers in the WM training literature (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; more than 1,000 citations!) concludes that training transferred to enhanced intelligence based on two observations. First, subgroups that trained with WM tasks showed more gains than passive controls. Specifically, transfer was compared to no practice at all, which could perhaps be attributed to a placebo effect. Second, four subgroups were trained with a different number of training sessions. The benefits on the untrained intelligence task differed across these subgroups, and were monotonically larger for subgroups that had a larger number of training sessions. Jaeggi et al.'s (2008) interpretation was that longer training periods induced a larger amount of improvement and transfer.

This interpretation seems compelling. However, the different subgroups were trained at different sites. A valid design would either be to conduct the entire experiment at one site, or divide each site into subgroups with different amounts of training per site, so that training sites and amount of training would not be confounded. These sites may thus have differed in terms of the experimenters' and/or participants' motivation. If so, the different amounts of practice as well as the different extents of transfer could reflect different degrees of enthusiasm with respect to the training procedure. In fact, when the experiment was replicated without these confounds, Chooi and Thompson (2012) obtained null results. The observation that different training sites (or different labs) yield different amounts of improvement for the same training procedure is rather typical (e.g., Jaeggi and colleagues consistently reported transfer—Jaeggi et al., 2008; Jaeggi, Buschkuehl, Jonides, & Shah, 2011; Jaeggi, Buschkuehl, Shah, & Jonides, 2014; Klingberg et al., 2005—versus

Redick et al., 2013, who used an active control and reported no transfer). Therefore, the importance of this dissociation is not only theoretical. Furthermore, the intuitive interpretation that longer training protocols yield larger amounts of improvement and transfer, though seemingly straightforward, is not supported by a recent meta-analysis (Au et al., 2014). It showed that across training studies, there was no correlation between the total amount of practice and the transferred gains. Similarly there was no correlation between the total gain on the trained task and the amount of transfer. This tends to weaken the authors' interpretations and it can only be concluded that the Jaeggi et al. (2008) study does not provide clear evidence of transfer from training on WM to enhanced intelligence.

This same meta-analysis of WM training studies (Au et al., 2014) identified a key difference between the transfer reported by studies that only had a passive control and studies that had an active control; namely, the former reported a significantly larger transfer than the latter (detailed in Jacoby & Ahissar, 2013). In fact, studies that used an active control group that practiced a challenging non-WM task (e.g., visual search) found no transfer (e.g., Owen et al., 2010; Redick et al., 2013; see also the meta-analysis by Melby-Lervåg & Hulme, 2013 and reviews by Morrison & Chein, 2011 and Shipstead et al., 2012).

The case of control in perceptual learning

Traditionally, perceptual learning studies had no control group. Their main goal was to show specificity; hence control was not essential. However, when the emphasis is on transfer effects, a valid control is required to assess the mechanisms underlying generalization. The importance of such a control has been somewhat overlooked in perceptual studies, perhaps due to the assumption that performance in perceptual tasks is mainly determined by peripheral sensory mechanisms.

In fact, an earlier training study conducted by our group suffered from this flaw. Banai and Ahissar (2009) trained a group of individuals with language and reading difficulties on a series of auditory discrimination tasks. Our participants' pretraining performance was very poor on these tasks compared to their adequately reading peers, and so was their verbal WM score. After several weeks of training, their performance on the trained perceptual tasks reached the level of their untrained, adequately reading peers. A posttraining test on their verbal WM skills showed that it also improved and reached the level of their (untrained) peers. We interpreted these observations as indicating transfer from the perceptual training procedure to verbal WM. But this could have reflected more vigilant posttraining performance stemming from general aspects of our training protocol. It may have induced more rewarding experiences in school, either due to personal attention, or due to our small tokens of appreciation for participation. Moreover, given that we had no passive control group with similar pretraining difficulties, it may have reflected a retest effect, particularly since this standard test, like many other standard tests, had only one version, so we used the same items.

Another type of inadequate control is a group that does not train (passive control), and hence controls only for test retest effects. For example, Deveau and colleagues (Deveau et al., 2014a, b) administered a visual training protocol composed as a video game designed to enhance general visual skills. In both studies, posttraining performance was better than pretraining performance on a variety of tasks, including visual acuity (measured with self-paced standard eye charts), contrast sensitivity, and peripheral acuity. Moreover, posttraining performance exceeded that of passive controls. However, as in our earlier auditory studies, the improved posttraining performance could have been due to factors that did not stem directly from their specific perceptual training regime, such as supportive meetings with enthusiastic experimenters.

Importantly, performance even on low-level tasks aimed to assess acuity and contrast sensitivity is affected by nonsensory factors. For example, when participants are primed (with a mindset) that pilots have excellent vision, their performance on basic visual tasks improves when they are asked to take the role of pilots (by flying a realistic flight simulator; Langer, Djikic, Pirson, Madenci, & Donohue, 2010). Similarly, action video gamers, and individuals trained with action video games for several weeks, perform better on contrast sensitivity tasks (Li, Polat, Makous, & Bavelier, 2009). This improvement has been attributed to improved top-down control of task-related information (Green & Bavelier, 2012). Recent observations nevertheless suggest that this enhancement requires some task-specific training, and that gamers' superiority on simple visual tasks reflects better top-down control rather than modified sensory mechanisms (Bejjanki et al., 2014).

The impact of state of mind on performance on simple perceptual tasks suggests that even active control groups may not constitute a sufficient control since participants' expectations can significantly affect their performance. For this reason, the active control group should train on a similarly challenging task. Perceptual learning researchers' traditional intuition that a control group should be presented with similar stimuli (e.g., Polat, Ma-Neim, Belkin, & Sagi, 2004) rather than with a similarly engaging task fails to control for this type of general effect.

However, even when the control group plays a challenging game, their more specific expectations, as well as those of the experimenters, may impact their posttest performance. For example, studies by Green and Bavelier (2003, 2006a, 2006b, 2007, 2012) and Green, Li, and Bavelier (2010) reported that individuals who practiced action video games performed better than those who practiced nonaction games (e.g., *Tetris*, *Sims*). Therefore, in these studies the active control also had a challenging game. Nevertheless, these findings have been contested in several studies conducted in labs with different expectations (e.g., Boot, Blakely, & Simons, 2011; Boot, Kramer, Simons, Fabiani, & Gratton, 2008; Lee et al., 2012; Van Ravenzwaaij, Boekel, Forstmann, Ratcliff, & Wagenmaker, 2014). These researchers claimed that the difference in the outcomes obtained in different labs reflects the experimenters' bias, which affects participants' anticipation of improvement, and consequently their performance (Boot, Simons, Stothart, & Stutts, 2013; see review in Kristjánsson, 2013). Importantly (as in the case of multisite training in the Jaeggi et al., 2008 paper), undocumented (and probably unintentional) differences between training protocols at different sites may have had a greater impact on the outcomes than the explicit choice of training procedure. These differences may be as influential in perceptual training as they are in training for more complex cognitive skills.

Correlation between training gains

A detailed analysis of a specific example of WM training

Since the magnitude of training generalization is often moderate (Jacoby & Ahissar, 2013; Melby-Lervåg & Hulme, 2013), it is natural to seek additional supportive evidence. One such line involves investigating the correlation between practice-induced gains on the trained task and performance enhancement on untrained tasks. This analysis seems compelling, but as shown below, significant correlations with a magnitude well within the reported range can be obtained with no transfer.

To clarify this point we present a detailed analysis of a specific case of WM training (Jaeggi et al., 2011), where the claim of generalization relies uniquely on the observation of such positive correlations. Based on the raw data (cordially provided by the authors), we suggest an alternative and coherent explanation for all of the experimental observations, assuming no transfer.

Jaeggi et al. (2011) trained two groups of third graders, one ($N = 32$) on general knowledge and vocabulary (active control), and the other ($N = 32$) on a spatial WM task. This initial study yielded null results. Namely, the two groups did not differ in their general intelligence scores, either before or after training, suggesting no transfer (using two standard test: Raven's matrices and Test of Nonverbal Intelligence; "there was no significant group × test-session interaction"; Jaeggi et al., 2011, p. 10081). However, rather than acknowledging null results, Jaeggi et al. (2011) argued that the correlation within the experimental group between gains on the trained WM task and the intelligence tests they administered was indicative of transfer. This argument may seem intuitive: Those who benefit from training on one task are expected to show gains on another task. However, such a correlation can be obtained without any transfer (Jacoby & Ahissar, 2013; Tidwell, Dougherty, Chrabaszcz, Thomas, & Mendoza, 2014).

Consider the scenario of pre- and posttraining measurements. Each participant has four scores: the pre- and posttraining scores on trained task A, $X_{1,A}$ and $X_{2,A}$, respectively, and the pre- and posttraining scores on test task B, $X_{1,B}$ and $X_{2,B}$, respectively.

The correlation between training gains (post minus pre) in these two tasks, $\Delta_A = X_{2,A} - X_{1,A}$ and $\Delta_B = X_{2,B} - X_{1,B}$, can be formulated as follows:

$$\mathrm{Corr}(\Delta_A, \Delta_B) = \frac{\mathrm{cov}(\Delta_A, \Delta_B)}{\mathrm{std}(\Delta_A)\mathrm{std}(\Delta_B)} \tag{1}$$

where:

$$\begin{aligned}\mathrm{cov}(\Delta_A, \Delta_B) = {} & \mathrm{cov}(X_{2,A}, X_{2,B}) + \mathrm{cov}(X_{1,A}, X_{1,B}) \\ & - \mathrm{cov}(X_{2,A}, X_{1,B}) - \mathrm{cov}(X_{1,A}, X_{2,B})\end{aligned} \tag{2}$$

Therefore the correlation in gains can be positive regardless of the means of the populations.

In the case of Jaeggi et al. (2011) and based on the authors' original data, the covariance between scores on the two tasks within sessions was larger than (session 2) or similar to (session 1) the covariance across sessions, yielding a positive covariance between gains in the two tasks.

In order to understand whether the specific covariance matrix obtained by Jaeggi et al. can reflect other factors rather than transfer, we assumed the following general factors:

- A common factor that contributes to the performance of two tasks that are initially correlated, for example in the case of WM and intelligence tasks (see Daneman & Carpenter, 1980).
- Common factors that contribute to the performance of tasks assessed within the same session ("good" vs. "bad" day).
- Common factors that contribute to the performance of the same task across sessions (i.e., bottlenecks that are not resolved by a second assessment).

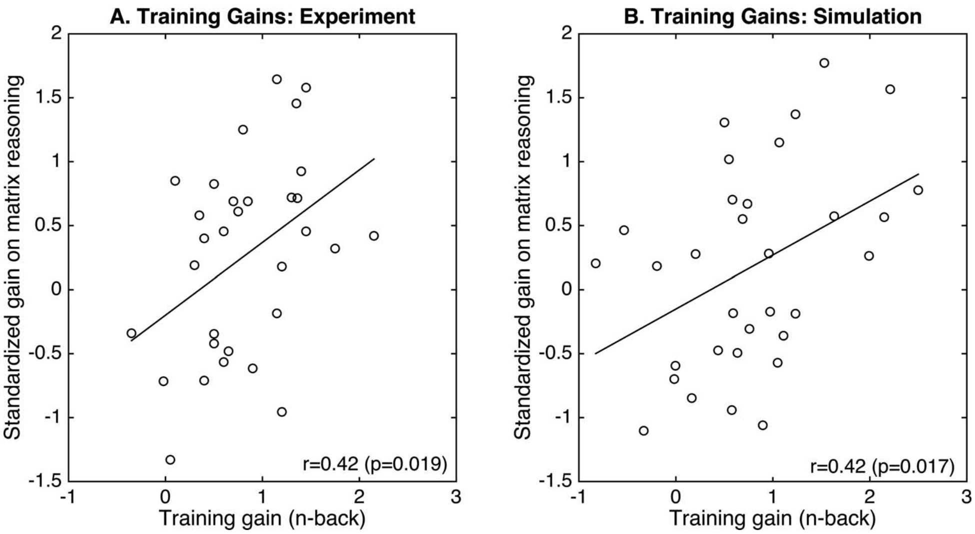

Figure 1. Gains in intelligence scores (matrix reasoning) as a function of gains on the trained WM task (*N*-back). (A) The original observations in Jaeggi et al. (2011; plotted based on personal communication with S. M. Jaeggi and M. Buschkuehl, and similar to the original supplemental figure 3sb). (B) One rendition of the simulated data, which was artificially designed to have no transfer. The similarity between the two plots demonstrates the similarity between the *r* value reported by Jaeggi et al. (2011) and that obtained by the simulation. The details of the simulation procedure are provided in the Appendix.

As shown in the Appendix, assuming that performance of each task reflects a linear combination of these five factors (commonality between tasks, two sessions, and two tasks), we can replicate the covariance matrix of Jaeggi et al., 2011, without assuming any transfer. Importantly, the calculated factor loadings are all positive and within the expected cognitive range (Engle et al., 1999b), as detailed in the Appendix.

Figures 1–3 demonstrate that the factors described above are sufficient to fully replicate all the results reported in Jaeggi et al. (2011). Figure 1 compares the observed correlation between gains in WM and in intelligence scores (left, their replotted supplemental figure S3B) with one rendition of a simulation that used their unpublished correlation structure (namely, with no difference in the mean performance of the subpopulations of high and low WM "gainers," respectively). This experimentally observed correlation was interpreted in Jaeggi et al. (2011) as "suggesting that the greater the training gain, the greater the transfer" (p. 10083). However, this correlation ($r = 0.42$, $p < 0.05$) is similar to the one obtained in our Monte-Carlo simulation. Quantitatively, 37.6% of the simulations (out of 10,000 renditions) had a larger (more significant) r value than the one reported in their paper, thus indicating that the significant correlation found in the experiment is consistent with the simulated data, which did not assume transfer.

Another analysis proposed by Jaeggi et al. (2011) has a similar conceptual flaw. They divided the group trained on the WM task into two subgroups (median split) as a function of high and low gains on the trained WM task. They found that the subgroup that had larger WM gains also had larger intelligence gains, and interpreted this observation as another indication of a transfer effect. As explained in Tidwell et al. (2014), this result is completely explained by the same principles.

Intuitively, this a-posteriori criterion biases the assignment of individuals to the two groups in the following manner. Individuals who have large differences between pre- and posttraining performance on the trained task ("large gain" subgroup) tend to be individuals who have lower pretraining scores and higher posttraining scores on this task. These individuals also tend to have lower pretraining scores on the untrained task due to the pretraining correlation between the trained and untrained tasks. In fact, the simulated F statistics (the main ANOVA statistics used in the paper to assess the transfer effect) were larger than the reported F statistics, $F(2, 58) = 3.23$, in 38.6% of the 10,000 simulations. This means that the values reported in the study lie well within the expected range of our simulation. Furthermore, using the planned contrast method, Jaeggi et al. (2011) reported a significant difference between the subgroups with large and small training gains ($p < 0.05$), and no significant differences between the active control group and the subgroup with small training gains. Our simulations (with the same simulation parameters as above) also reproduced this pattern: In 47% of the renditions, the first contrast was significant ($p < 0.05$) and the second contrast was not ($0.05 < p < 0.95$). Figure 2 shows the pre- and post-WM scores in the experiment (left two plots) and in the simulation (right two plots). The similarity of the plots stems from the a-posteriori

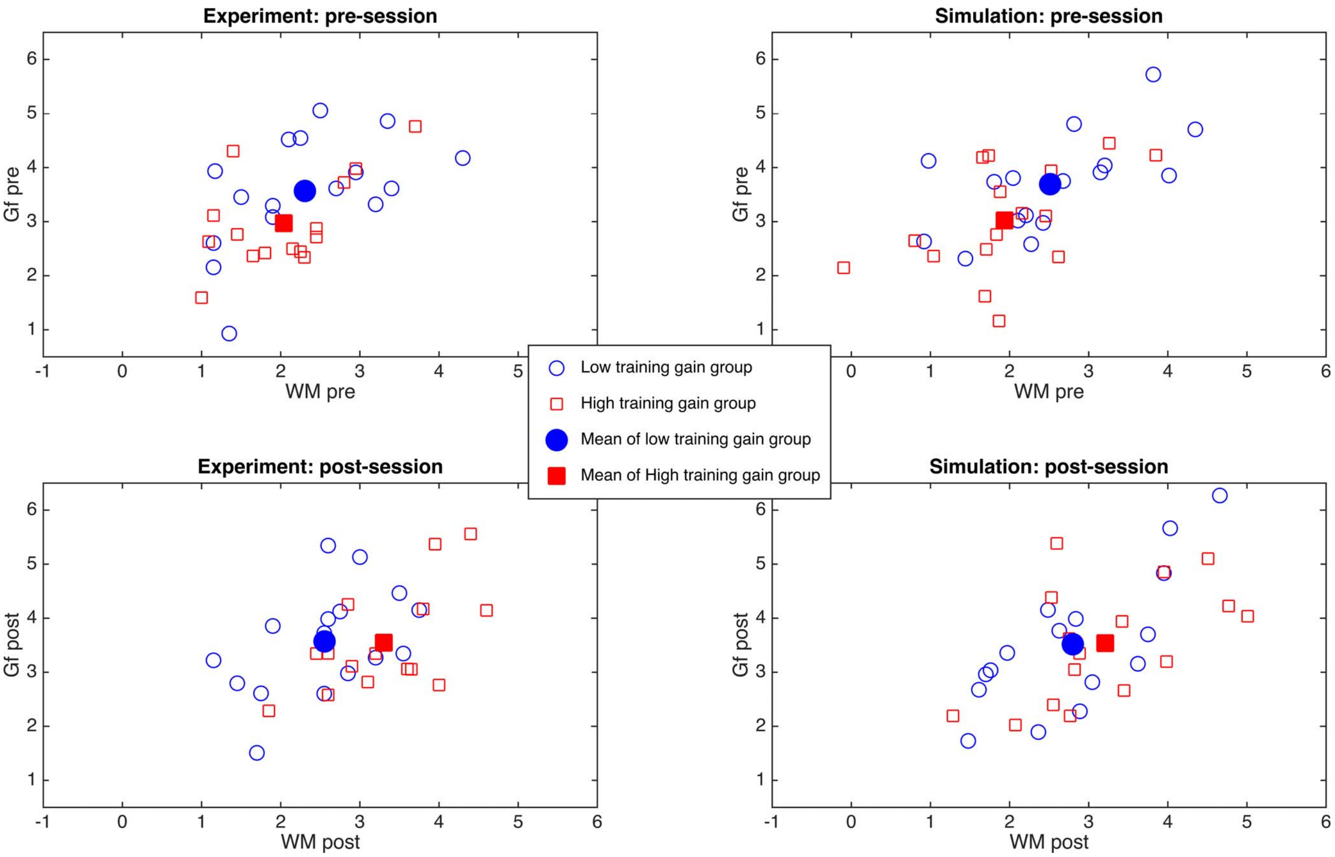

Figure 2. Pre (top) and post (bottom) WM and intelligence scores in the experiment (left) and in one rendition of the simulation (right). The open symbols indicate the scores of the individuals composing the two a-posteriori selected subgroups (blue circles for low training gain and red squares for high training gain). Note that individuals with low WM gains tend to have relatively high WM scores pre training and relatively low WM scores post training. Filled symbols denote the means of these subgroups. The plots illustrate the similarity between the experimental results and the simulation.

division into subgroups rather than from a transfer effect.

Jaeggi et al. (2011) further reported a larger transfer for the subgroup with larger training gains than the group with smaller training gains (and for the intermediate transfer of the active control group). These observations are plotted in the top left of Figure 3 (reported in figure 4a of Jaeggi et al., 2011). Figure 3 top right shows the transfer effects in our simulation. The simulated data were generated so that all scores were chosen from a single distribution of performance, but were subsequently divided into two subgroups with the posthoc criterion of the original paper (below and above median gains on the trained task). The similarity of the right and left top plots of Figure 3 indicates that the difference between the subgroups is fully reproduced in the simulated data.

Jaeggi et al. (2011) retested the participants several months after training to evaluate the long-term retention of the transfer effects. These measures suggested the retention of this effect, as shown in Figure 3 bottom left. But, as shown in the bottom right of Figure 3, this long-term retention was also fully reproduced by the simulation by using this posthoc selection criterion. Finally, we reproduced the original ANCOVA test used to attempt to control for numerical differences between the pretest scores of participants with the large and small training gains: 81.4% of our simulations had a larger (more significant) F value than the one reported in the paper (3.06) on a univariate ANCOVA with the mean standardized gain as the dependent variable and the pretest scores as the covariate. Additionally, 6.6% of our simulations had a smaller p value on this test than the one reported in the paper ($p < 0.01$) for the follow-up session.

To conclude, the authors should have accepted their own null findings based on their a-priori design. Correlations in gains should only have been used to claim transfer when the magnitude of the experimental effects was larger than that which could be obtained based on the a-posteriori selection criterion. As our analysis showed, this was not the case.

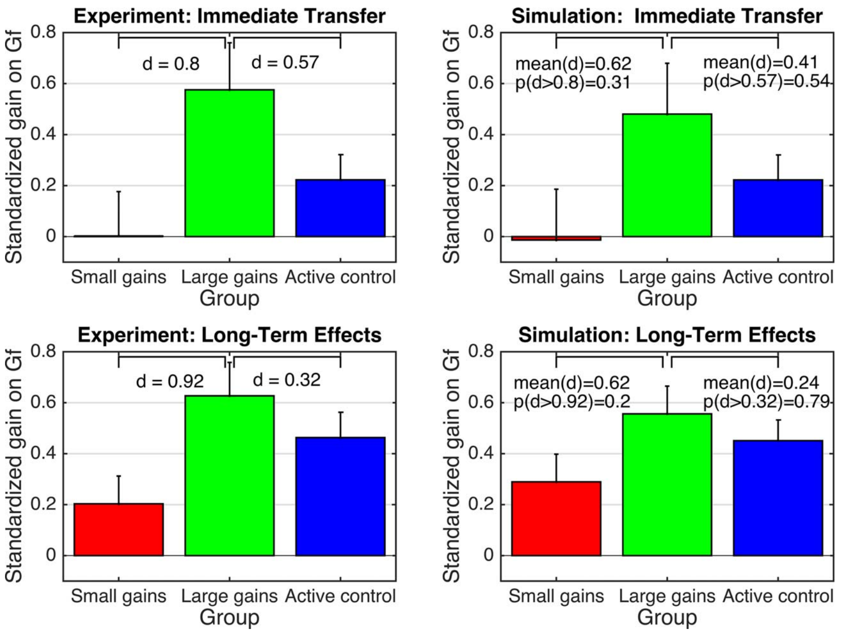

Figure 3. Immediate (top) and long-term (bottom) gains in the (fluid) intelligence tests (Gf) of the two experimental subgroups determined a-posteriori by the magnitude of their training gains (denoted in red and green) and the active control group (denoted in blue). Left: The experimental data, replotted (figure 4a, b of Jaeggi et al., 2011). Right: The results of the simulation. Mean transfer (Cohen's *d*) for 10,000 renditions of our simulation are plotted for each subgroup. No transfer was assumed in the simulation. The posthoc selection criterion was sufficient to yield "transfer" results. Columns represent standardized gains. The effect size of the contrast between the subgroups was computed with Cohen's *d*. The effect sizes of the empirical data are consistent with those yielded by the simulations. For example, 31% of the Cohen's *d* computed by the simulation for the large-gain versus small-gain subgroups were larger than the experiment's *d* (0.80).

We next examined what would be the expected pattern of results in Jaeggi et al. (2011) had the correlation in gains indeed reflected a genuine transfer. First, we would expect that the two subgroups (determined by WM gains) would differ in their posttraining intelligence scores. Those that had a higher WM gain would be expected to have higher post training intelligence scores, which was not the case in the original study (where the two subgroups differed mainly in their pre-training scores). Second, we would expect that the entire group's posttraining intelligence scores would be more variable than their pretraining scores, since training with WM had a different effect (transfer) on different individuals. This was not the case either. Third, if transfer is assumed in a simulation (transfer was simulated as 1 *SD* between means of the subgroups), the Cohen's *d* of the difference in intelligence gains between the subgroups should have been larger than the reported one in 99.7% of the simulations. The mean Cohen's *d* of the simulations in this case was 1.8, whereas the experimental one was only 0.8. Therefore, if real transfer were involved, we would expect a larger degree of estimated transfer than the one reported by Jaeggi et al. (2011).

Additional examples from the WM literature

Jaeggi et al. (2011) is one of many examples of using correlated performance benefits as an indication of transfer in the literature of WM training. For example, Chein and Morrison (2010) tested participants on a cognitive battery that included fluid intelligence, reasoning tasks, reading comprehension, cognitive control, and WM tasks. They then trained these participants for 4 weeks on two variants of WM tasks. This group was subsequently tested with the same cognitive battery as in the pretraining session. A passive control group (21 students) was administered the test–retest battery without training. The authors claimed to have found "a strong and statistically significant relationship between trained participants' spatial WM span increases and reading comprehension improvement, $r(18) = 0.49$, $p < 0.005$" (p. 197). The authors further claimed that their transfer effect was selective, in the sense that there was no similar correlation in the control group, $r(20) = 0.097$, $p = 0.67$. As explained above, this observation can be fully accounted for on the basis of Equation 1, even in the absence of transfer. This study also demonstrates the other problem (subdivision according to gains). They divided their training group into successful (the 15 participants with higher spatial-WM gains in the training group) and unsuccessful subgroups (the rest of the group). When they compared the successful group to the control group, they found marginally significant and significant transfer effects for the cognitive control and reading comprehension tasks, respectively.

A similar analysis was reported by Klingberg, Forssberg, and Westerberg (2002), who trained children with ADHD in an attempt to enhance intelligence scores, as measured by Raven's matrix completion task. They claimed that "the association between the reasoning task and the WM tasks is further substantiated by the significant correlation between improvement on the visuospatial WM task and improvement on Raven's Progressive Matrices" (p. 789). Similarly, Schmiedek, Lövdén, and Lindenberger (2010) noted a correlation between latent trained factors of episodic memory and untrained ones, and used it as evidence of transfer.

Similar cases in perceptual learning and the video gaming literature

Similarly flawed interpretations have been in the literature on perceptual learning. The aforementioned study conducted by our group (Banai & Ahissar, 2009), likewise suffered from this methodological flaw. As described above, we trained individuals with language and reading disabilities on auditory discrimination tasks. We tested their verbal WM skills before and after training and found improvement and observed gains in WM. The trained group showed a marginally significant correlation between improvement on the trained perceptual task (two-tone discrimination) and on an untrained WM task (Spearman rho $= 0.59$, $p = 0.07$). We (Banai & Ahissar, 2009) interpreted this correlation as supporting the suggestion that the improvement in WM scores is specifically related to the two-tone discrimination training.

This intuitive interpretation is quite common in the literature of video game training. For example, Green and Bavelier (2003) used the marginal correlation between improvement in game scores and improvement in attentional skills (adjusted $r^2 = 0.43$, $p = 0.13$; see p. 536) to support a transfer claim between playing action video games and enhanced attentional skills. More recently, Anguera et al. (2013) trained elderly individuals on a dual-task condition of an action video game and used the correlation between improvement in this trained condition (multitasking) and WM gains to support the claim of transfer from training multitasking in action video games to enhanced WM capacity. They reported that "only the multi-tasking [training] group exhibited a significant correlation between multitasking cost reduction and improvements on an untrained cognitive control task (delayed-recognition with distraction) from pre- to post-training" (p. 99). This reasoning presents the very same methodological flaw (i.e., assuming that only the dual task was initially correlated with WM, as would be expected given the cognitive literature; Clapp, Rubens, Sabharwal, & Gazzaley, 2011; Kane & Engle, 2000).

Note that methodological flaw does not only apply to the correlation in behavioral gains. It also applies to misinterpreting correlations between behavioral gains and the magnitude of changes in related brain measures that have the same characteristics where the pretraining magnitude is correlated with the measured behavior. These correlations are similarly sensitive to the common factors described in the section titled "A detailed analysis of a specific example of WM training." However, they are often used as indicating training induced modulations of the underlying brain mechanisms. For example, Wu et al. (2012) trained participants on a first-person shooter game (FPS) and tested them on an attentional visual field task (AVF). The control group played a puzzle game. Event-related potentials (ERPs) were measured during the AVF test before and after training. Similar to Jaeggi et al. (2011) the authors divided the FPS group into high (FPS+) and low (FPS–) performers according to their improvement in accuracy on the AVF test, and compared the mean gains in amplitudes of ERP components (P2 and P3). Changes in the ERP components were only observed in the FPS+ group, and overall there was a main effect for group (the groups were FPS+, FPS–, and control). The authors interpreted this finding as demonstrating "a direct causal relationship between playing an FPS video game and the neural activity that supports spatial selective attention" (Wu et al., 2012, p. 1290). However, this interpretation suffers from the same methodological problems. Pretraining correlations together with the posthoc splitting of the training group according to training gains can again account for the results, even in the absence of a causal relation. We should note that we assume pretraining correlations between ERPs and behavioral accuracy, which were measured simultaneously, but the correlations were not reported. Given this assumption, figure 3 in their paper, which shows differences in changes in neural activity between the control group and the two sub groups defined based on training gains (FPS+ and FPS–), is completely analogous to figure 3 in Jaeggi et al. (2011), which shows transfer gains in similarly selected subgroups (defined based on posthoc training gains). In both cases the more parsimonious interpretation is the one that does not imply a causal relation, as our simulations demonstrate (see our Figure 3).

Conclusion

In this paper we discussed two methodological issues in the WM training literature: the lack of a valid control group with similar challenges and characteristics, and the overinterpretation of correlated gains between training and transfer tasks. We analyzed two studies in depth (Jaeggi et al., 2008; Jaeggi et al., 2011) to highlight general methodological difficulties that undermine many other studies in the domains of WM, video games, and perceptual learning. In particular, training studies tend to fortify partial findings by a-posteriori selection of supportive evidence.

The aim of these analyses is not to discourage further studies in the field of training, or the assessment of potential transfer. Rather it stresses the need to analyze nonintuitive methodological areas, which should be carefully addressed in future studies to achieve a better understanding of the actual factors that determine generalization.

Keywords: perceptual learning, perceptual training, working memory, learning transfer, generalization

Acknowledgments

This work was supported by the Israel Science Foundation (grant 616/11), the EPFL-HUJI Brain Collaboration and the Gatsby Charitable Foundation. Nori Jacoby is supported by the ELSC Brain Sciences Postdoctoral Fellowship. We are grateful to Susanne Jaeggi and Martin Buschkuehl for providing the data from their article and for commenting on the first draft of this paper. N. Jacoby would like to thank Carmel Raz for her help during the preparation of the manuscript.

Commercial relationships: none.
Corresponding author: Nori Jacoby.
Email: jacoby@mit.edu.
Address: Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences, Cambridge, MA, USA.

References

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences*, *90*(12), 5718–5722.

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, *387*(6631), 401–406.

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, *8*(10), 457–464.

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1515), 285–299.

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

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2014). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *22*(2), 366–377.

Banai, K., & Ahissar, M. (2009). Perceptual learning as a tool for boosting working memory among individuals with reading and learning disability. *Learning & Perception*, *1*(1), 115–134.

Bejjanki, V. R., Zhang, R., Li, R., Pouget, A., Green, C. S., Lu, Z. L., & Bavelier, D. (2014). Action video game play facilitates the development of better perceptual templates. *Proceedings of the National Academy of Sciences*, *111*(47), 16961–16966.

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

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*(3), 387–398.

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). 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.

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

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

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

Chooi, W. T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*(6), 531–542.

Clapp, W. C., Rubens, M. T., Sabharwal, J., & Gazzaley, A. (2011). Deficit in switching between functional brain networks underlies the impact of multitasking on working memory in older adults. *Proceedings of the National Academy of Sciences*, *108*(17), 7212–7217.

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Science*, *7*(12), 547–552.

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

Deveau, J., Lovcik, G., & Seitz, A. R. (2014a). Broad-based visual benefits from training with an integrated perceptual-learning video game. *Vision Research*, *99,* 134–140.

Deveau, J., Ozer, D. J., & Seitz, A. R. (2014b). Improved vision and on-field performance in baseball through perceptual learning. *Current Biology*, *24*(4), R146–R147.

Dosher, B. A., & Lu, Z. L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences*, *95*(23), 13988–13993.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999a). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Mechanisms of Active Maintenance and Executive Control* (pp. 102–134). New York: Cambridge University Press.

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

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

Fiorentini, A., & Berardi, N. (1981). Learning in grating waveform discrimination: Specificity for orientation and spatial frequency. *Vision Research*, *21*(7), 1149–1158.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*(2), 171–191.

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

Green, C. S., & Bavelier, D. (2006a). Effect of action video games on the spatial distribution of visuospatial attention. *Journal of Experimental Psychology: Human Perception & Performance*, *32*(6), 1465–1478. doi:10.1037/0096-1523.32.6.1465

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

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

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

Green, C. S., Li, R., & Bavelier, D. (2010). Perceptual learning during action video game playing. *Topics in Cognitive Science*, *2*, 202–216.

Green, C. S., Pouget, A., & Bavelier, D. (2010). Improved probabilistic inference as a general learning mechanism with action video games. *Current Biology*, *20*(17), 1573–1579.

Herzog, M. H., & Fahle, M. (1997). The role of feedback in learning a vernier discrimination task. *Vision Research*, *37*(15), 2133–2141.

Jacoby, N., & Ahissar, M. (2013). What does it take to show that a cognitive training procedure is useful? A critical evaluation. *Progress in Brain Research*, *207*, 121–140.

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*(25), 10081–10086.

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & Cognition*, *42*(3), 464–480.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*(2), 336–358.

Karni, A., & Sagi, D. (1991). Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences*, *88*(11), 4966–4970.

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

Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal of Clinical and Experimental Neuropsychology*, *24*(6), 781–791.

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

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence*, *14*(4), 389–433.

Langer, E., Djikic, M., Pirson, M., Madenci, A., & Donohue, R. (2010). Believing is seeing: Using mindlessness (mindfully) to improve visual acuity. *Psychological Science*, *21*(5), 661–666.

Lee, H., Boot, W. R., Basak, C., Voss, M. W., Prakash, R. S., Neider, M., … Gratton, G. 2012. Performance gains from directed training do not transfer to untrained tasks. *Acta Psychologica*, *139*, 146–158.

Levi, D. M., & Polat, U. (1996). Neural plasticity in adults with amblyopia. *Proceedings of the National Academy of Sciences*, *93*(13), 6830–6834.

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.

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

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

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

Polat, U., Ma-Naim, T., Belkin, M., & Sagi, D. (2004). Improving vision in adult amblyopia by perceptual learning. *Proceedings of the National Academy of Sciences, USA*, *101*(17), 6692–6697.

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

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods* (Vol. 319). Citeseer. New York: Springer-Verlag.

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience*, *27*(2), doi:10.3389/fnagi.2010.00027.

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

Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S. S., & Merzenich, M. M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science*, *271*(5245), 81–84.

Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J. R., Thomas, R. P., & Mendoza, J. L. (2014). What counts as evidence for working memory training? Problems with correlated gains and dichotomization. *Psychonomic Bulletin & Review*, *21*(3), 620–628.

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*, 1286–1293.

van Ravenzwaaij, D., Boekel, W., Forstmann, B. U., Ratcliff, R., & Wagenmakers, E. J. (2014). Action video games do not improve the speed of information processing in simple perceptual tasks. *Journal of Experimental Psychology: General*, *143*(5), 1794–1805.

Appendix

Details of the Monte-Carlo simulations

We created multiple renditions of populations of simulated participants and modeled the performance of each participant in every rendition as a multivariate Gaussian distribution (the same means as the empirical data and the same covariance matrix). We then repeated the posthoc selection process described in Jaeggi et al., 2011 for each rendition, by splitting into subgroups based on the training gains. We show that this a-posteriori division replicated all the statistical results reported in Jaeggi et al. (2011) by comparing the empirical statistical results with the expected results from multiple renditions of the simulation.

In this simulation, each participant's score is represented as a five-dimensional random variable $X = (X_{1,W}, X_{1,G}, X_{2,W}, X_{2,G}, X_{3,G})$.

$X_{1,W}$ and $X_{2,W}$ are scalar random variables which represent pre-WM and post-WM scores, respectively. $X_{1,G}$, $X_{2,G}$, and $X_{3,G}$ are random variables that represent standardized pretraining, posttraining, and follow-up measures of fluid intelligence, respectively. These scores combine Raven's Standard Progressive Matrices and the Test of Nonverbal Intelligence. The vector with the simulated scores of each subject (X) was sampled from the same multivariate Gaussian distribution $X \sim N(\mu,\Sigma)$. The mean (μ) and covariance matrix (Σ) were taken from the average of the entire WM training group (data provided by S. M. Jaeggi & M. Buschkuehl, personal communication, October 2013) and were:

$$\mu = (\mu_{1,W}, \mu_{1,G}, \mu_{2,W}, \mu_{2,G}, \mu_{3,G}) \\ = (2.10, 3.24, 2.92, 3.50, 3.64).$$

Other than a normalization constant and a minor difference in the treating of missing data, these are identical to Jaeggi et al., 2011; Table 1, second row.

The covariance matrix was calculated according to the correlation matrix and the empirical standard deviation for each task, as presented in Table A1.

For the active control group, each participant's scores were given by a similar five-dimensional vector $X^{AC} = (X_{1,C}, X_{1,G}, X_{2,C}, X_{2,G,}, X_{3,G})$. Again, X^{AC} was randomized from the distribution $N(\mu^{AC}, \Sigma^{AC})$ where the mean $\mu_{AC} = (0.59, 3.29, 0.58, 3.51, 3.74)$ was computed from empirical data. The covariance matrix was the empirical covariance matrix of the control group (see Table A2).

Having established the structure of the simulation, we simulated the statistical posthoc selection criterion of the original paper by splitting the group trained with WM into two subgroups with large and small WM gains (below and above the median of $\Delta_W = X_{2,W} - X_{1,W}$). The results, presented in the main text (Figures 1 and 2), show the outcome of this a-posteriori selection criterion (i.e., that it was sufficient to generate the results in Jaeggi et al. (2011) even though we introduced no transfer).

In order to quantitatively compare the reported statistical results to our simulations, we used the Monte-Carlo method (Robert & Casella, 2004) and randomized multiple renditions of the simulation. We computed the statistical tests for each rendition, and compared them with the ones reported in the original paper. We made sure that the statistical tests reported in the paper did not lie in the lower or upper percentile of the simulated statistics (nonsignificant).

	W1	G1	W2	G2	G3
W1		0.62	0.74	0.62	0.65
G1			0.41	0.65	0.93
W2				0.63	0.56
G2					0.76
SD	0.765	0.930	0.772	0.812	1.216

Table A1. Correlation coefficients and standard deviations in the training group. W1 and W2 are the WM scores in sessions one and two, respectively. G1, G2, and G3 are the intelligence scores in sessions one, two, and the follow-up session.

	C1	G1	C2	G2	G3
C1		0.73	0.38	0.64	0.81
G1			0.27	0.79	0.83
C2				0.20	0.29
G2					0.78
SD	0.016	0.646	0.031	0.733	0.827

Table A2. Correlation coefficients and standard deviations in the active control group. C1 and C2 are the scores in the active control task in sessions one and two, respectively. G1, G2, and G3 are the intelligence scores in sessions one, two, and the follow-up session.

In order to show that the correlation structure itself is not related to transfer, we show below that a positive covariance in gains, and even a replication of the entire covariance matrix with all assessed tasks are compatible with a no-transfer assumption. The same covariance matrix can be produced with no transfer, assuming that performance on the assessed tasks is determined by a linear combination of the following factors:

- U_0 is the common factor that contributes to the performance of the demanding WM and intelligence tasks (see Daneman & Carpenter, 1980).
- U_1 and U_2 are common factors the contribute to the performance in tasks assessed within the same session. Since correlation between tasks may vary across sessions, we used U_1 and U_2 for the common performance within Sessions 1 and 2, respectively.
- U_W and U_G are test–retest commonality factors for WM and for fluid intelligence, respectively. These factors link scores of the same test measured in different sessions.

The definitions of these factors do not necessarily indicate correlation (i.e., they are not necessarily positive values) but they simply allow for this eventuality option. Let us consider a scenario where the performance vector X is determined in the following way:

$$\begin{bmatrix} X_{1,W} \\ X_{1,G} \\ X_{2,W} \\ X_{2,G} \\ X_{3,G} \end{bmatrix} = \begin{bmatrix} g_1 & a_1 & 0 & c_1 & 0 \\ g_2 & a_2 & 0 & 0 & d_1 \\ g_3 & 0 & b_1 & c_2 & 0 \\ g_4 & 0 & b_2 & 0 & d_2 \\ g_5 & 0 & 0 & 0 & d_3 \end{bmatrix} \begin{bmatrix} U_0 \\ U_1 \\ U_2 \\ U_W \\ U_G \end{bmatrix} + \mu \tag{A1}$$

where g_1, g_2, g_3, g_4, g_5 are affinity constants determining the contribution of the shared WM and intelligence factor, U_0, to performance in each of the five measures composing the Vector X; a_1, a_2, b_1, b_2 are the affinity constants that determine the contribution of the sessions' factors to performance; c_1, c_2, d_1, d_2, and d_3 are the affinity constants that determine the contribution of task-specific factors to performance, and μ is the empirical mean of the scores.

We assume that all the factors in Equation 3 are standardized independent Gaussian random variables, and choose $g_1 = 0.76$; $g_2 = 0.52$; $g_3 = 0.66$; $g_4 = 0.65$; $g_5 = 0.82$; $a_1 = 0.41$; $a_2 = 0.29$; $b_1 = 0.13$; $b_2 = 0.55$; $c_1 = 0.11$; $c_2 = 0.56$; $d_1 = 0.76$; $d_2 = 0.30$; $d_3 = 0.74$.

The constants were fitted using numerical optimization in order to obtain a covariance matrix of the random variables described by Equation A1 that is almost identical to the empirical covariance matrix (Table A1). Note that the obtained factors are all positive and in the expected range (Engle et al., 1999b). Since all of the simulation results in this paper were based on the empirical covariance matrix, we thus show that all the statistical results reported in Jaeggi et al. (2011) could be obtained without transfer.